LADES User's manual

Alan Vázquez-Alcocer

Centro de investigación en Matemáticas, A.C.

Copyright © 2014 Alan Vázquez-Alcocer

HTTP://CIMAT.MX/ HECTORHDEZ/LADES/INDEX.HTML

The LaTeXtemplate used for this manual is the 'Legrand Orange Book' originally created by Mathias Legrand. This material is used in compliance with the Licence of the Creative Commons Attribution-NonCommercial 3.0 Unported License, `http://creativecommons.org/licenses/by-nc/3.0`.

First printing, February 2014

Contents

1 Introduction 5

1.1 LADES Help! 6

1.2 How to install LADES? 6

1.3 User Sessions 6

1.4 Import and Export 6

2 Create Module 7

2.1 Full Factorial Design 7

2.2 Fractional Factorial Designs with 2-levels 12

2.3 Optimal Design of Experiments 16

2.4 Full Longitudinal Design 21

2.5 Unequal Longitudinal Design 24

2.6 Longitudinal Optimal Powered-Cost Design 29

3 Evaluate Module 35

3.1 Longitudinal Design 35

3.1.1 Summary of Evaluating Longitudinal Designs 35

3.1.2 Evaluating a Longitudinal Design 35

3.1.3 Function Outputs 37

4 Analyze Module 39

4.1 Design of Experiments Statistical Analysis 39

4.2 Longitudinal Analysis 50

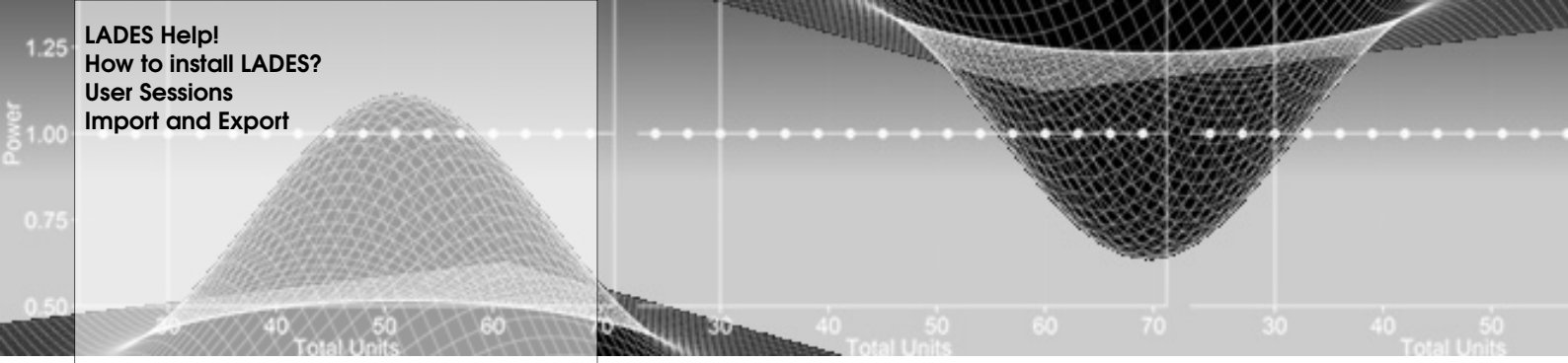
4.2.1 Linear Mixed Model 50

4.2.2 Generalized Linear Mixed Models 68
4.2.3 Generalized Estimating Equations 76
4.2.4 Generalized Least Squares, GLS 82

5 Statistics Module 93
5.1 Linear Regression 93
5.2 Logistic Regression 103
5.3 Sample Size Estimation for Longitudinal Analysis with Continuous Response 110
5.4 Sample Size Estimation for Longitudinal Analysis with Binary Response 113

6 Graphics Module 117
6.1 Longitudinal Data Graph 117
6.1.1 Function outputs 119

A Term Definitions 125
A.1 Model Matrix 125
A.2 Contrasts 125
A.3 Variance Structures 126
A.4 Correlation Structures 127
A.5 Likelihood Criterias 128



1 — Introduction

Welcome to LADES Software v. 2.0.0 (Longitudinal Analysis and Design of Experiments Software). LADES's main goal is to provide an accessible platform using the most advanced tools for the analysis of experimental data. Its design contains statistical methods for:

- Design and Analysis of Experiments
- Longitudinal Data Analysis
- Sample Size Estimation
- Visual Elements for Analysing Longitudinal Data

LADES supplies researchers and practitioners with a user-friendly interface with independent working sessions. Furthermore, it has the capability to export and import *.csv* files, which makes your results easier to share and evaluate. LADES is intended to include all the necessary tools when analyzing experimental results (laboratory, industrial, chemical, etc)

All functions of LADES are based on the R language for statistical computing and graphics. LADES makes use of R libraries such as:

- MASS
- lattice
- nlme
- lme4
- FrF2
- AlgDesign
- geepack
- ggplot2

LADES was built using JAVA programming language and it is available in its project home page: `http://cimat.mx/~hectorhdez/lades/index.html`. LADES' requirements are:

- OS: Windows XP, 7 and 8
- JAVA 1.6 or superior

1.1 LADES Help!

This document is divided into six Chapters that will help you to get a better understanding of the capacity and power of LADES. First of all, we show how to install LADES in your computer and we talk about user's sessions, input and output, Chapter 1. Chapter 4, presents tools for creating a variety of designs; Chapter 5 shows a function to evaluate longitudinal designs; Chapter 6 explain all the functionality regarding the analysis of the resulting data; Chapter 7 presents a variety of statistical methods with different purposes such as fitting model to compute sample size; Chapter 8 shows the longitudinal data graph function. Finally, an Appendix is included to provide definitions of some important concepts used throughout the manual.

For each function an introduction, a summary of all important concepts and theory regarding it, and an example implemented in LADES are provided.

1.2 How to install LADES?

Before you can install LADES, you have to update or install JAVA software. You can download the last version of JAVA from `http://java.com/es/download/index.jsp`. Once you have downloaded it, please just follow the installer instructions.

If you want to install LADES on your computer, please follow:

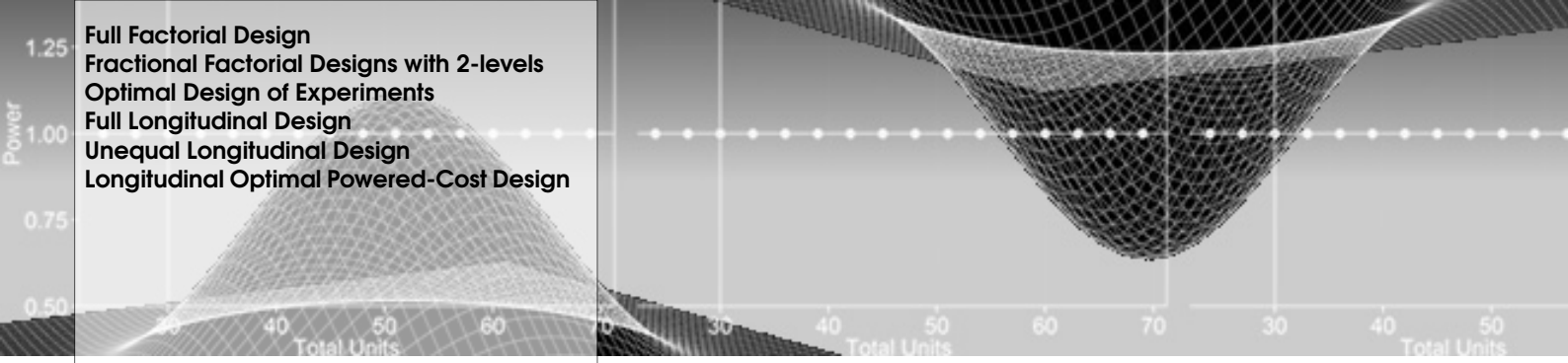
1. Download the **LADES.exe** file from `http://cimat.mx/~hectorhdez/lades/donwload.html`
2. Open the **.exe** file and follow the instructions. The **R 2.15.1** installer will appear automatically; just follow the instructions of it.
3. Open LADES by double-clicking its desktop icon.
4. The username to start your first LADES session is: USERNAME: **admin** and PASSWORD **1234567**. You may be asked to install the R libraries, just click YES.
5. Start using lades and have fun!

1.3 User Sessions

If you want to start a new session in LADES go to the main menu and click on **File** $->$ **Administration** $->$ **Users Control**. If you want to include a new user just click on **new**.Just fill the required information and click **Save**. Now you can start a new session under your user.

1.4 Import and Export

To import a .csv file to LADES just click on **Import** and choose the file from your computer. If you want to save a file just click on **Save**. To add or delete columns from your data click on the bottoms **+** or **-** on the bottom of the main menu.



2 — Create Module

2.1 Full Factorial Design

Most of the experiments are designed to analyze the effects of two or more factors as well as possible interactions between them. In general, full factorial designs are the most efficient for this type of experiments. By factorial design we mean that on each run of the experiment, all combinations of factor levels are investigated. For example, see Table 2.1 where a full factorial design to analyze the effects of diets, including Fibre and Fat as factors, is shown. For Fibre factor, researchers wanted to study the effect of a high fibre with a low content. For Fat, three levels were proposed.

Run	Fibre	Fat
1	Low	Low
2	High	Low
3	Low	Medium
4	High	Medium
5	Low	High
6	High	High

Table 2.1: Experimental Design for Diet data

The runs are each of the combinations of the two factors, or each possible diet. By using this design, only six runs are required.

Summary of Full Factorial Designs

- The most popular designs of experiments are the full factorial designs. The 2-level full factorial design is referred as 2^k, where k is the number of main factors. This design consists of 2^k runs (the total number of combinations). Two important properties the 2^k factorial designs have are *balance* and *ortogonality*. Balanced means that each level of each factor appears in the same number of runs. Two factors are said to be orthogonal if all combinations levels appear in the same number of runs. A design is called orthogonal if all pairs of factors are orthogonal. These two properties are very important because they allow us to have more accurate estimates of the effects of each of the main factors and interactions, as well as to have a clearer interpretation of the results.

- The contrasts used for the design refer to encodings for each level of each factor. For example, if we have Temperature as a two level factor with 10^oC y 30^oC, we can encode this factor to be -1 for 10^oC and 1 30^oC. These new encoded levels help us to make our design orthogonal as well as the model matrix. This lead us to obtain estimates of the effects for each factor. Now, If we want to establish three levels for temperature, one way to encode our levels is: -1 for 10^oC, 0 20^oC and 1 30^oC. This type of Orthogonal contrasts are called **Ortogonal** or **Helmert**. On the other hand, if we have a factor whose levels are categories, such as high, medium and low; we would use other kind of contrasts called **Treatment** (see A.2). Such encodings would be Level 1 for low, Level 2 for medium and Level 3 for High. The model matrix for this design is constructed so as to be orthogonal. For categorical factors this construction of the orthogonal matrix is performed by decomposing each of the corresponding factors to their corresponding sub-factors level. Finally, the contrast selection (and model matrix) depends on the questions we want to answer.
- For a design whose factors have more than two levels, it is difficult to calculate and interpret their effects when they are categorical. The simplest way to calculate the effects of the factors is through the least squares method using the model matrix. This model matrix is calculated from the proposed design. For the case of 2^k designs, this matrix is always orthogonal, but in cases where we have more than two levels on each factors, the matrix undergoes some complications as interpretation of the results. The least squares method led us to obtain unbiased estimates for each of the factor effects when our design is orthogonal.
- The experiment can be replicated, i.e. all runs of a design were performed more than once. It is said a design that is not replicated have only one replica. Having more replicas of our experiment allows us to have more information; and thus, to perform better statistical analysis.

Constructing a Full Factorial Design

As an example, we will create the design mentioned at the beginning of this section using LADES. To access the function for creating full factorial designs we follow this path: **Create** $->$ **Full Factorial Design**. The main screen of this function is shown in Figure 2.1. We can add more factors to the design using the **>** button on the right upper corner of the screen. In the same manner, we can delete factors using the **<** button. To edit the names of the factors as well as their levels, we use the section labeled in green.

In Figure 2.2 we see the captured data for our example. We must note we used **Treatment** as our contrasts (see A.2) due to the fact we have a categorical factor with more than two levels.

Now, we will create another full factorial design for a study in which we want to analyze the effect a new treatment has on the number of red blood cells (RBC). The performance of the new treatment will be compared against the performance a placebo (control). We want to analyze these effects on two different mouse strains, the BALB / c and C57BL. Finally, we want to perform the whole experiment 3 times in order to obtain three replicas.

Then, to create this design, we access again to the **Create Full Factorial** function and we introduce two factors, adding their names as **Treatment** and **Strain**, and the levels for each of them. In this case there were only two levels for both factors (see Figure2.3).

Function Outputs

- The resulting design for the first example is shown in Figure 2.4. Levels are represented using the numbers 1 to 3. We can export the table using the **Export Table** button. After exporting the table, we can change the names levels for their true labels to conduct the

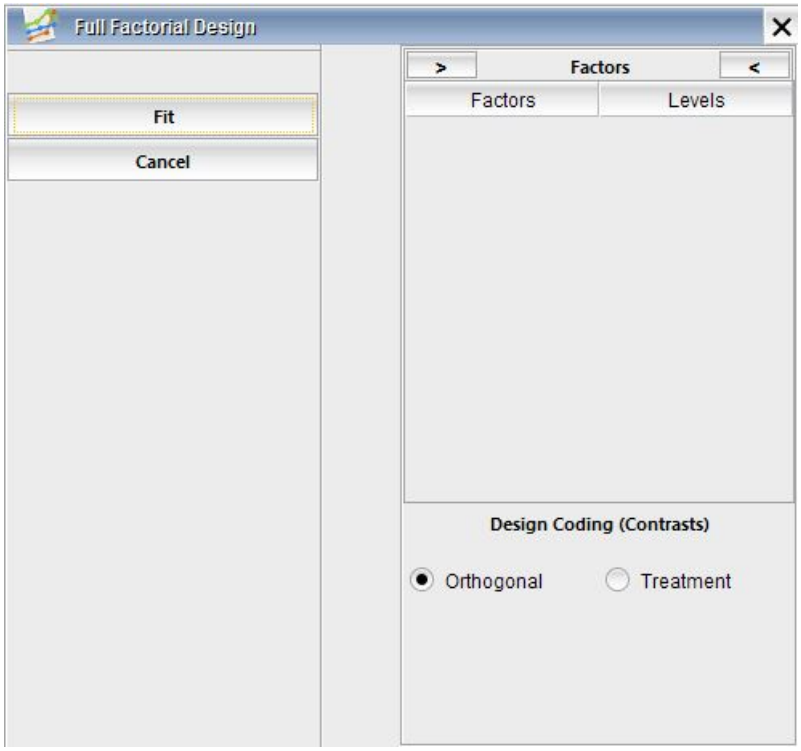


Figure 2.1: Display of the Function Create Full Factorial Designs

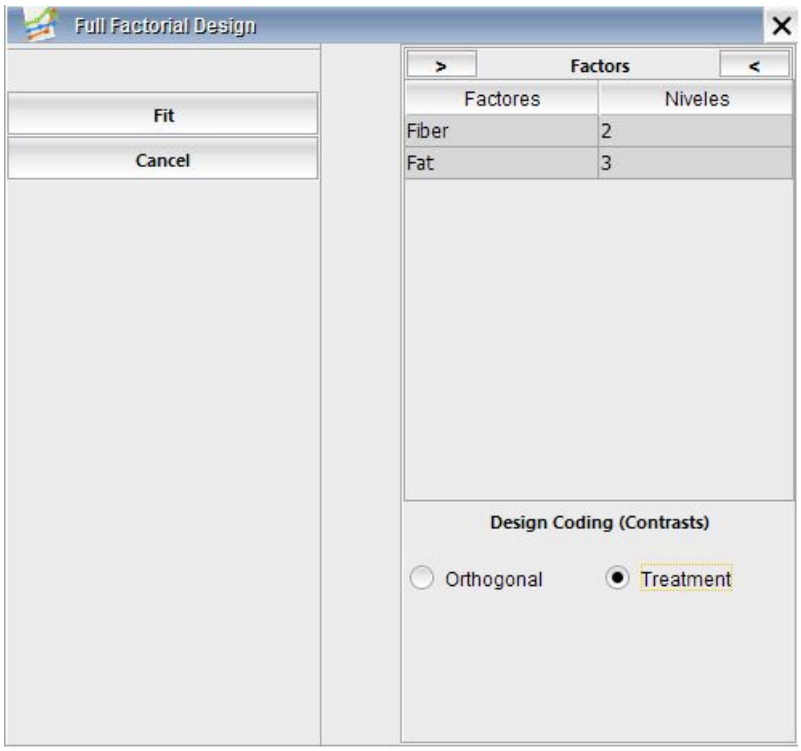


Figure 2.2: Create a full factorial design for the diets' experiment

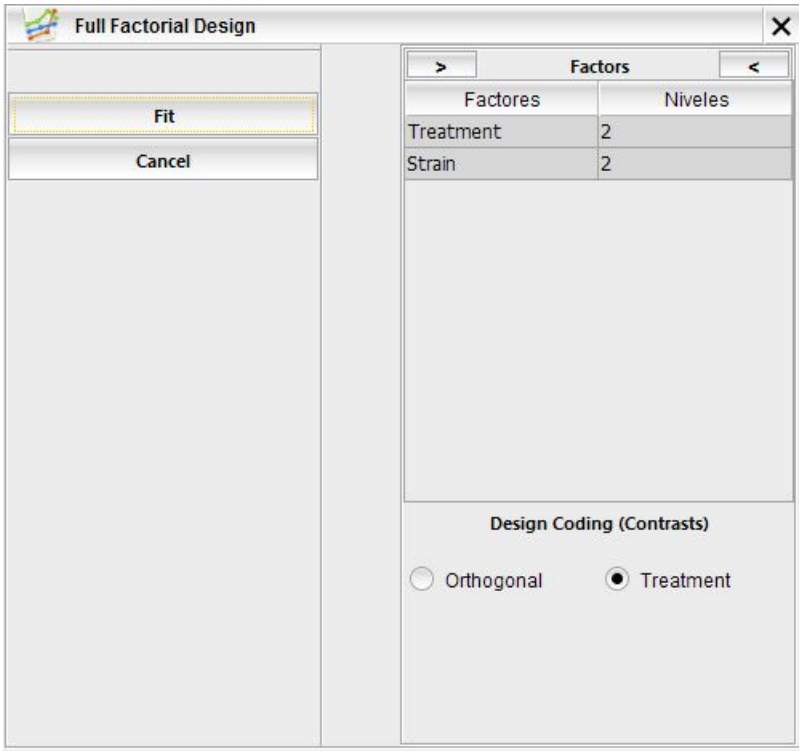


Figure 2.3: Create a full factorial design for the RBC's experiment

experiment. We can also randomize the runs of the experiment to meet some assumptions that we will discuss later. If we want to perform more replicas of the experiment we only have to copy the design several times and randomize.

- The resulting design for example 2 is shown in Figure 2.5. They used orthogonal contrasts (see Figure 2.3) due to the fact we have two factors with two levels each. This will have benefits later, when we do the analysis of the resulting data. We can also export and manipulate the design using other tools such as Excel. We can copy it twice to obtain the three replicas and then randomize the runs to meet the assumptions required by the statistical analysis' assumptions.

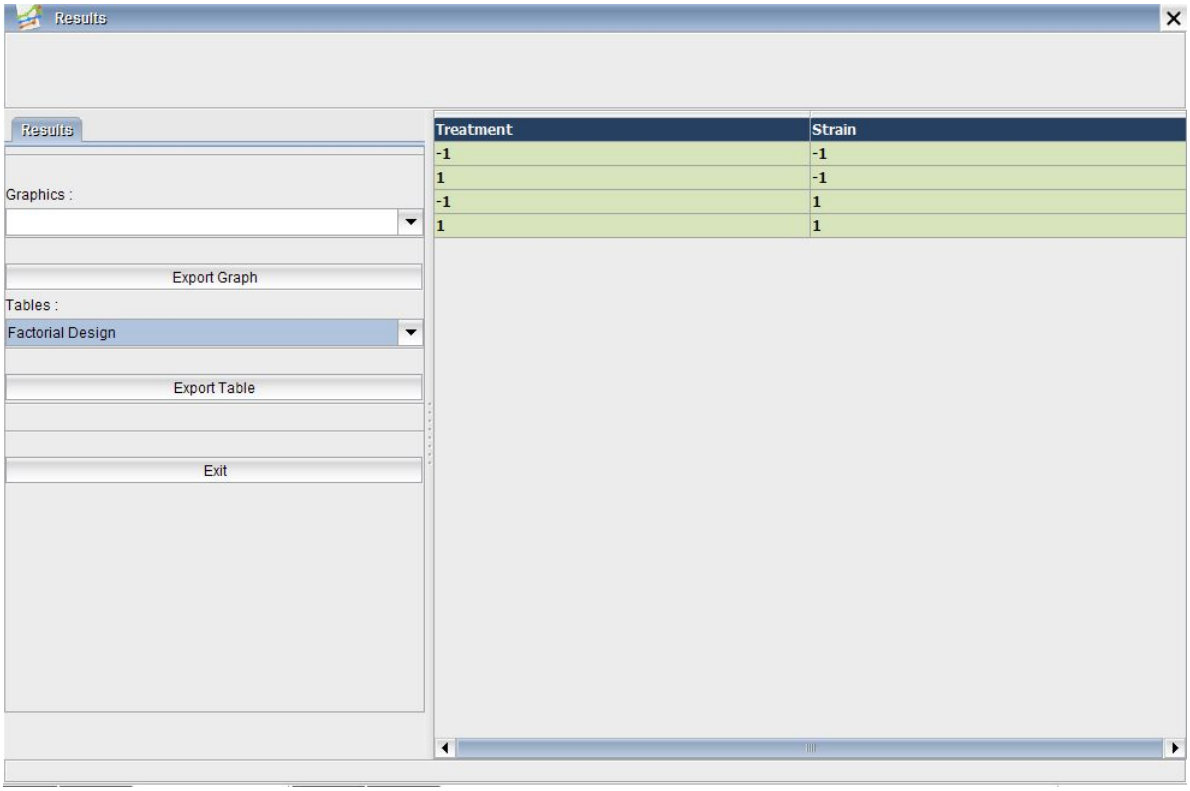


Figure 2.4: Full factorial design for the diets' experiment

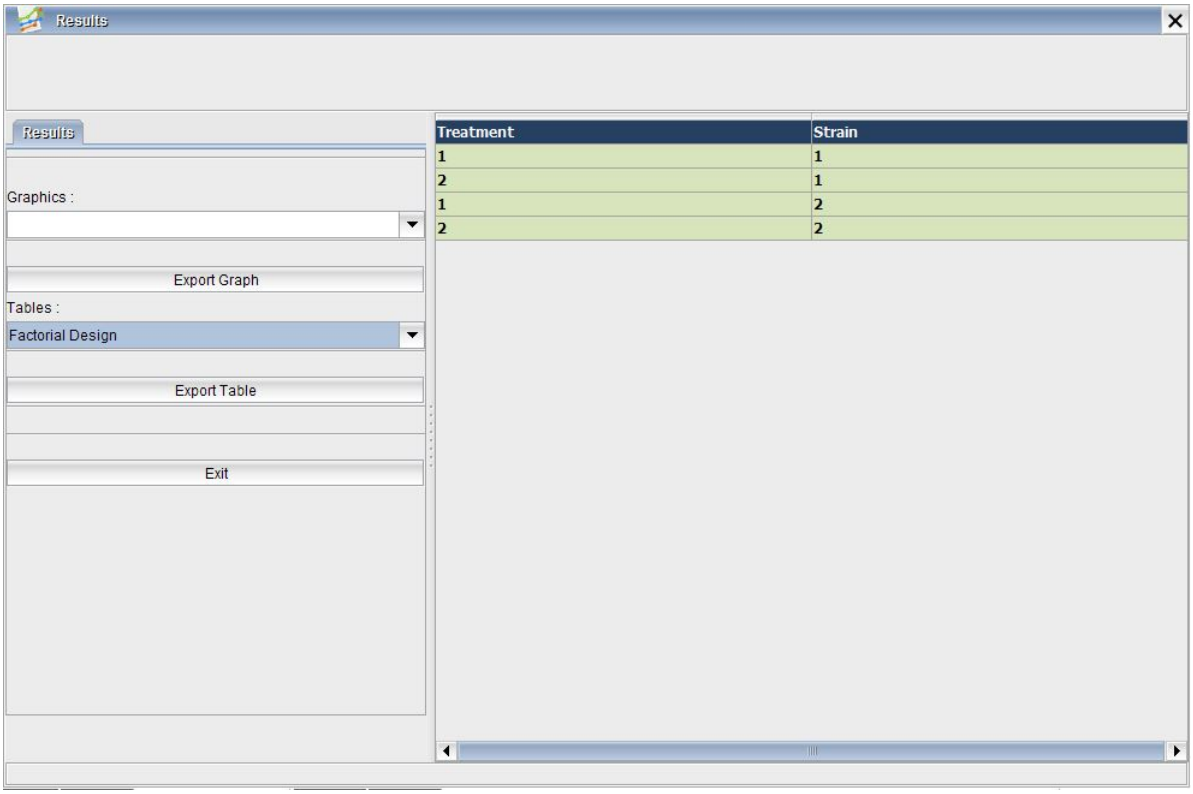


Figure 2.5: Full factorial design for the RBC's experiment

2.2 Fractional Factorial Designs with 2-levels

If the number of factors in a 2^k factorial design increases; then, the number of runs necessary for completing a replica quickly increases. For example, a complete replica of a design 2^6 require 64 runs. This seems to be too much regarding that only few effects, such as main effects or some interactions interactions, will be significant.

If we can reasonably assume that certain high-order interactions will not be significant; then, by using only a fraction of the complete design, we could get enough information about the main effects and some interactions of two factors. These fractional factorial designs are the most used in experiments where you want to reduce the number of resources used.

The most common use of the fractional factorial designs is as *screening*. In this type of experiment many factors are considered, and it aims to identify factors (if any) with the largest effects. Screening experiments are used in the early stages of a study, when many initial factors are considered for the study but it is known jus a few will have a significant effect.

The success of fractional factorial designs is based on the following concepts

1. **Sparsity.** When there are many factors under study, and it is expected that only a few main effects and interactions will be significant for the response.
2. **Projection.** The resulting significant effects of a fractional factorial designs can be projected on larger designs.
3. **Sequential Experimentation.** It is possible to combine some runs from two or more fractional factorial designs to build a new factorial design. Such design would allow us to get betters estimates for main effects and interaction effects of interest.

Summary of Fractional Factorial Designs

- Consider a situation where three factors, at two levels each, are of interest. But, the experimenter can not perform $2^3 = 8$ combinations of all levels of the factors. The experimenter only has budget for perfuming only four runs. Thus, this suggests us to use a **half fraction** of a 2^3 design. Since this design requires only $2^{3-1} = 4$ combinations levels of factors. A fraction of a half of a 2^3 design is called a 2^{3-1} design. Similarly, we can define a quarter of a 2^5 factorial design. This design is known as 2^{5-2} design, and requires only $2^{5-2} = 2^3 = 8$ runs instead of 32 runs for the complete design.
- Now, taking the fractionated design 2^{3-1}; how we construct the design? For this, consider Table 2.2. This table shows the model matrix (see A.1) of a 2^{3-1} design. We note this matrix is equal to the model matrix of a 2^2 factorial with four runs. To create the 2^{3-1} design we simply assign the third factor to the third column of the model matrix, i.e. the interaction between factor A and B. We define the **generator** design 2^{3-1} as **C = AB**, and this means the effect of C is confused with the effect of the interaction AB.

			C
I	**A**	**B**	AB
+	-	-	+
+	+	-	-
+	-	+	-
+	+	+	+

Table 2.2: Model matrix for a 2^{3-1} design

The process to generate a 2^{5-2} design is described as follows. In this case we have a 2^{5-2} design with 8 runs, or a 2^3 design. In such design, two factors are now confused. In order to create the generators we need to assign the other two C factors and D to either second order interactions as AB, BC, AC, or to the interaction ABC. The way how we

assign such actors and create the generators affects considerably an important property called **resolution**. Thus, we should be careful when making such assignments. Finally, we have to mention that if the factor C in a 2^{3-1} design is significant; then, the effect of the interaction AB is also significant, this due to the confounding property. So, the next step to get a better estimate, separately from the AB effect estimate, of the main effect of factor C is to generate extra runs. This method is known as **foldover**.

- The design 2^{3-1} worked up above is called design of resolution III. In this design, the main factors are confounded with two-factor interactions. Usually, Roman numerals are used to define the types of resolution. The designs of resolution III, IV and V are the most important. The definitions of each of these designs are as follows. **Designs of Resolution III** are designs in which no main effect are confounded with other main effects; but, such main effects are confounded with some two-factor interactions. **Designs of Resolution IV** are designs in which no main effects are confused with other main effects or second order interactions; but, the second-order interactions are confused on each other. In **Design of Resolution V**, main effects and interactions are not confounded with any other main effect or interaction; but, two-factor interactions are confounded with some three-factor interactions. The type of resolution depends on the generators used to create the design. Then, a criterion for choosing generators where we can get the maximum resolution.
- For constructing and analyizing such designs, it is usually assumed that interactions of order three or more are not significant.

Constructing Fractional Factorial Designs

To access the function for constructing fractional factorial design, we must follow the route **Create** $->$ **Fractional Factorial**. By choosing this function, the screen shown in Figure 2.6 appears. This function shows a catalog containing maximum resolution fractional factorial designs. In such table we can see: the number of runs required for each design, the number of factors that can be analyzed with the design, the name of the design, its generators and its resolution. We just simply choose the desired option and click on **Create**. For generators, the numbers indicate factors, for instance, 1 is the first factor, 2 is the second etc. Thus, for example in option 3, the factor 4 is confused with the interaction of the first and second factor; and factor 5 is confused with the interaction of factor 1 and 3.

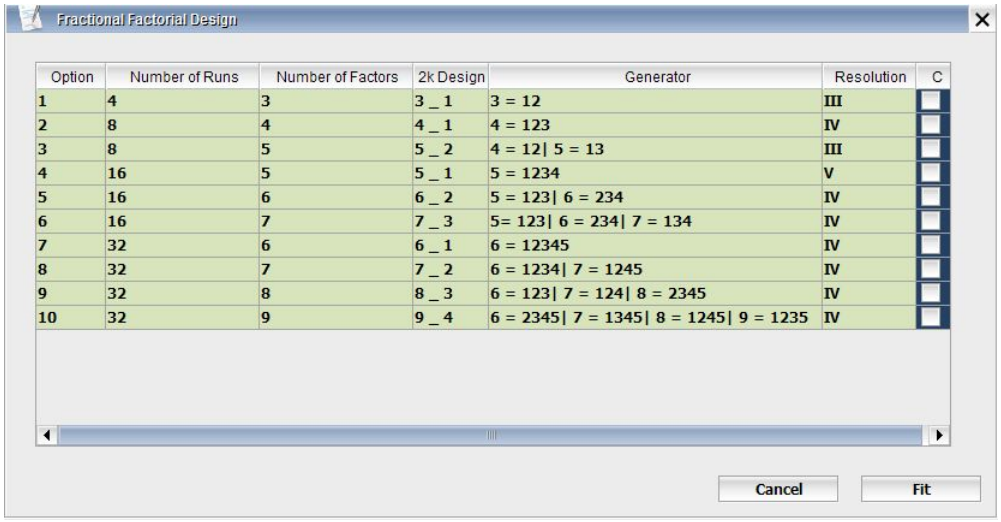


Figure 2.6: Main screen, Create Fractional Factorial

For example, if we want to create the 2^{5-2} for analyzing 5 factors, we only click on option 3 and then **Create**.

Function Outputs

- **Generators.** Displays the generators that were used to create the design. See Figure2.7

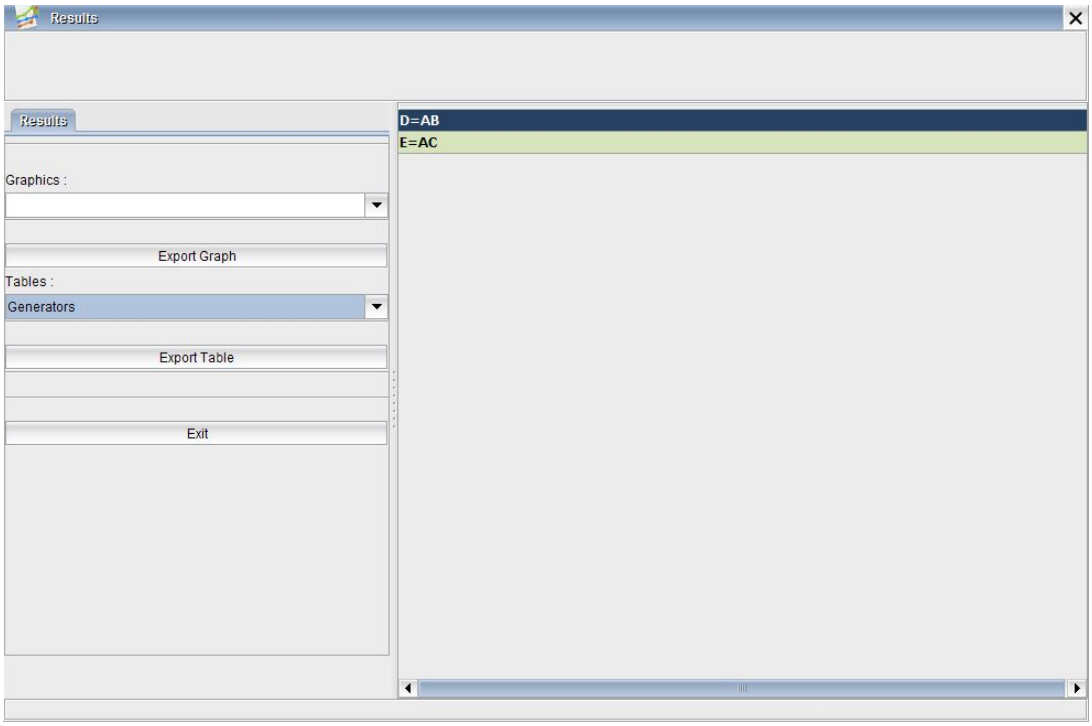


Figure 2.7: Generators, Create Fractional Factorial

- **Interactions.** Figure 2.8 displays the effect from which each two-factor interaction is confounded. For our example, only two interactions are confounded.
- **Main.** In this screen we can see the confounding for main effects. We note our main factors are confounded with two-factor interactions. Moreover, one of them is confused with more than one interaction. See Figure 2.9.
- **Fractional Factorial Design.** Figure 2.10 shows the generated fractional factorial design. This design can be exported to a *.csv* file by using the **Export Table** button.

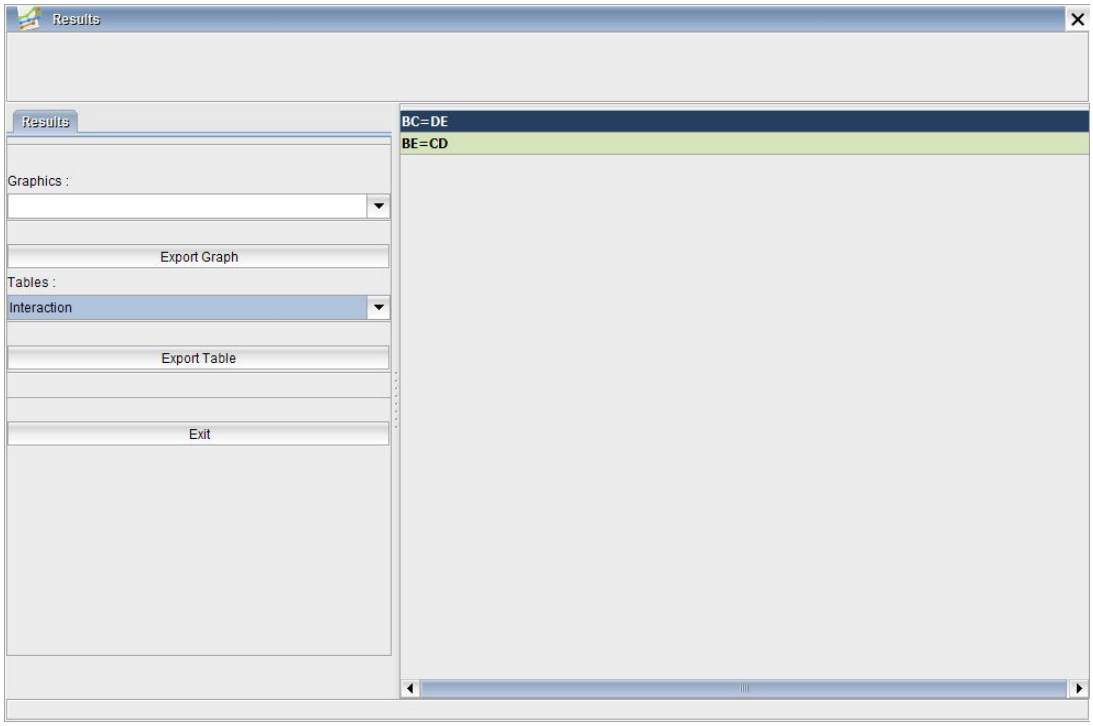


Figure 2.8: Interaction Confounding, Create Fractional Factorial

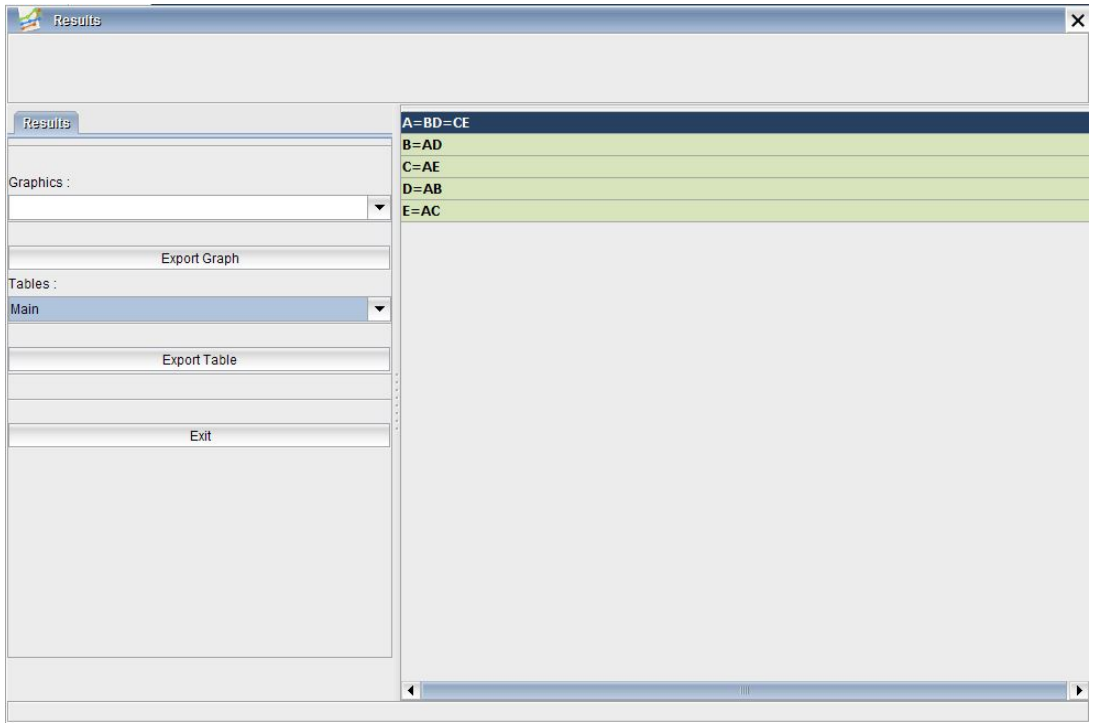


Figure 2.9: Main Effect Confounding, Create Fractional Factorial

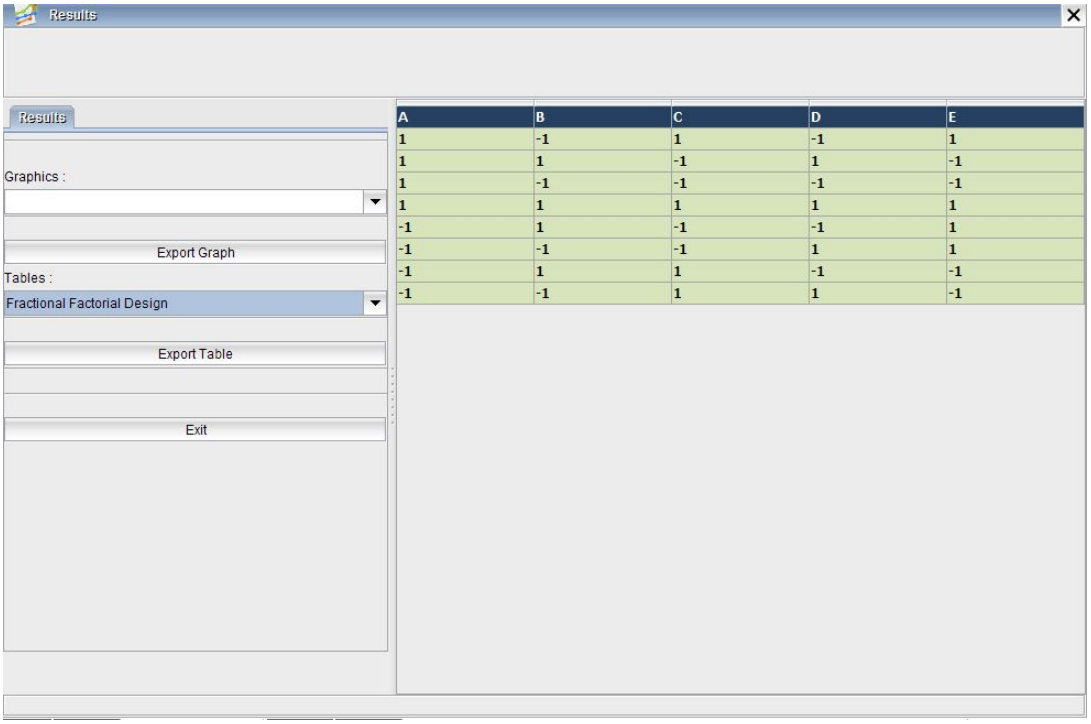


Figure 2.10: Fractional Factorial Design, Create Fractional Factorial

2.3 Optimal Design of Experiments

Usually, screening designs are two-level factorial designs used for fitting models with main effects and two-factor interactions. The benefit of considering only two factor levels and consider a simple model is the experiment can be perform using less runs. By using full factorial designs and fractional factorial design we have to choose over a large catalog of designs that best fits our goals. But, what happens if the available resources for the experiment did not allow us to use a full factorial design or a fractional designs? An example of this problem is when you only have resources to carry out only seven runs. The designs that we know are based on runs that are multiples of two, 2^k combinations, where k is the number of factors.

For this type of problem, we propose the best design according to our resources so that we can get more information from the observations. The area of statistics that works with this problem is called **optimal design of experiments**.

Summary of Optimal Design of Experiments

- As already mentioned, we must build the best design to explain as more information as possible. There are different criteria to classify the designs as better than others. The criteria used by this software is the D-optimality criterion. This criterion is based on giving the best estimates of main and interaction effects.
- One way to estimate the effects of main factors and interactions is through estimating the coefficients β_i of the regression model of interest (this may include only main effects, or main effects and some interactions). Coefficients are estimated using least squares, which leads to the the following notation:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

where X is the model matrix and Y is the vector of responses. The matrix covariance estimates of these coefficients is

$$Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

where X is the model matrix, and σ^2 is the variance of the model error. The D-optimality criterion is responsible for choosing the design that minimizes the determinant of the covariance matrix. This lead us to more accurate estimates for our coefficients. This minimization of the covariance matrix can be performed by maximizing the determinant $M = |X'X|$, since σ^2 is unknown but constant.

- The construction of these designs that minimize the variance of the estimates of coefficients is performed by using the point change algorithm.
- Full factorial and fractional factorial designs are ideal because of the ortogonality property; thus, they produce the minimum determinant for M (diagonal matrix). In order to measure the deviance from the perfect ortogonality of one optimal design we use the Variance Inflation Factor (VIF). The minimum value of VIF is one. For orthogonal designs, the VIF for each of the estimates effects of the factors is 1. When a design is not orthogonal, one VIF is greater than one. If the VIF of an effect is four, then variance of the estimate of such effect is four times greater than the estimate using an ortogonal design.

Constructing an Optimal Design of Experiments

To access the function for creating optimal designs we follow the next route **Create** $->$ **Optimal** 2^k **design**.

As an example, we have a study in which a new treatment to aid the healing process in rats is to be assessed. The population under study is two strains of rats, Wistar and NuNu, using males and females each. This new treatment is to be compared against a control group. Treatment was applied and the measurements were taken 3 days after application. It is considered that the interaction between the type strain and treatment can be significant. The response under study is area of the wound, measured in mm^2 . Only seven rats are available, so it is necessary to propose the best design to analyze the factors of interest.

In Figure 2.11 we see the main screen of the function. In this figure we can also see the construction of the model of interest

$$y = \beta_0 + \beta_1\, Strain + \beta_2\, Sex + \beta_3\, Treatment + \beta_4\, Strain : Treatment + \varepsilon \qquad (2.1)$$

In Figure 2.11 we input the factors that will be part of our model X_1, X_2, X_3 which are Strain, Sex and Treatment, respectively. We include the interaction of interest $X_1 : X_3$, i.e. Strain:Treatment. The size of the design is seven. And we will run the algorithm 100 times. Once data is inputed, we click **Fit**.

Function Outputs

- **Design Matrix Determinant.** This window displays the relative efficiency of the design. This is the ration of the $M_{newdesign}$ and $M_{FullFactorialDesign}$ (2^3). See Figure 2.12.
- **Optimal Design.** This section presents the best resulting design for model (2.3). See Figure 2.13.
- **Model Matrix.** In Figure 2.14 we show the model matrix used (2.3;also see A.1).
- **Variance Inflation Factor.** In Figure 2.15, we observe the VIF for main effects and interactions. The VIF for all is 1.16, this implies that the variance of the estimates of the

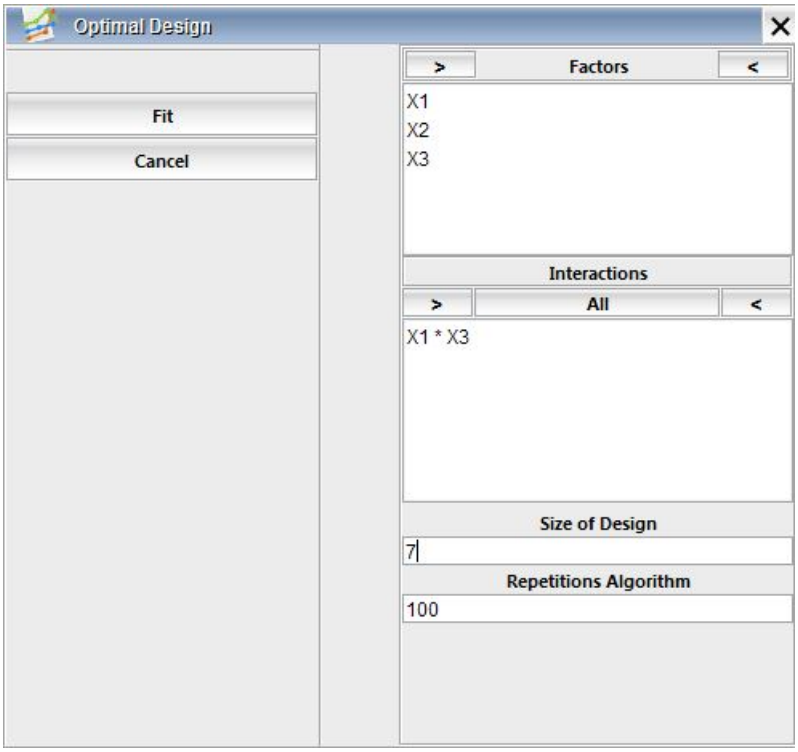


Figure 2.11: Create Optimal Design Screen

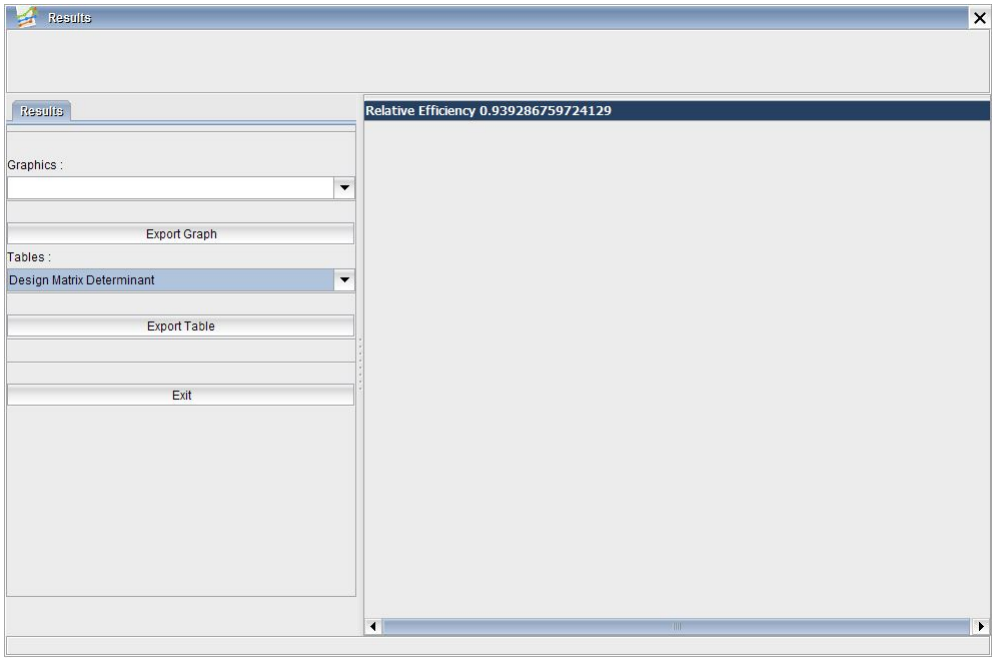


Figure 2.12: Relative Efficiency, Create Optimal Design

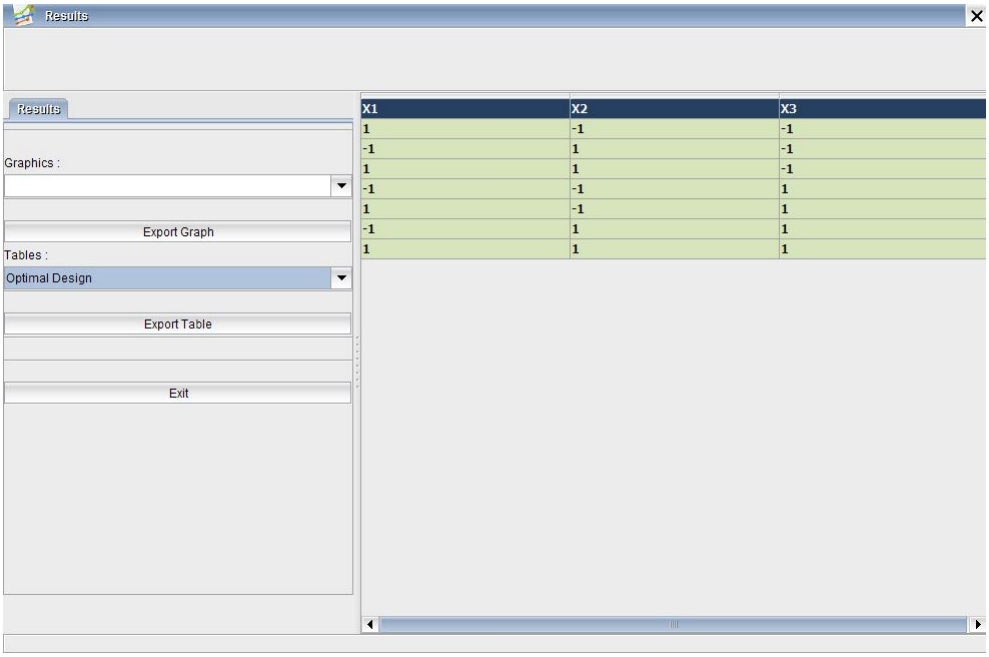


Figure 2.13: Optiml Design, Create Optimal Design

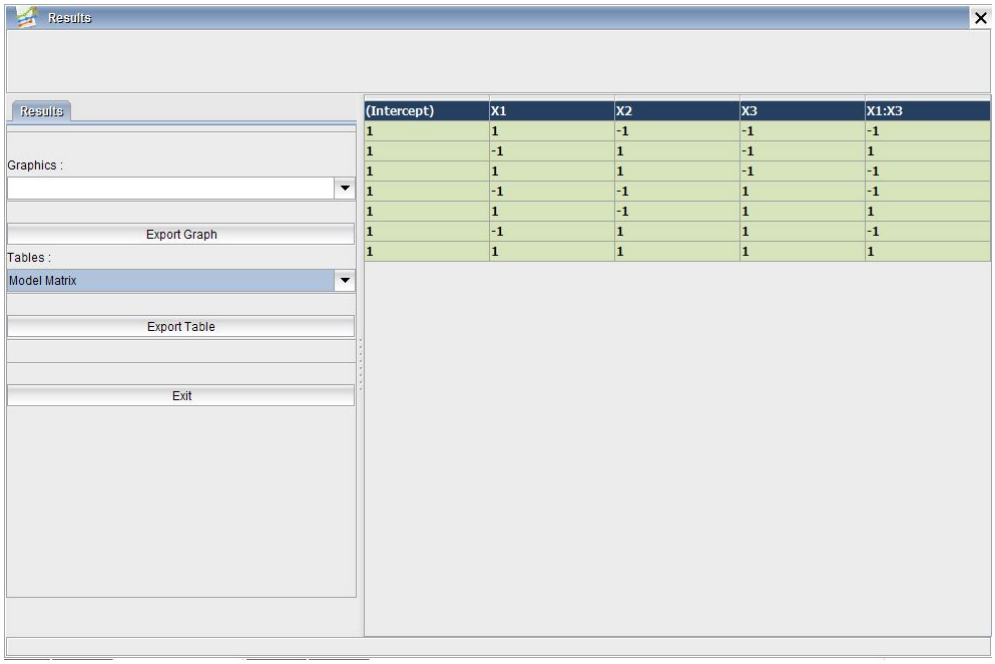


Figure 2.14: Model Matrix, Create Optimal Design

effects will be 1.16 larger than the estimated effect using an ortogonal design. Because that inflation is not very large, we accept the design. A rule of thumb is that VIF is not greater than five

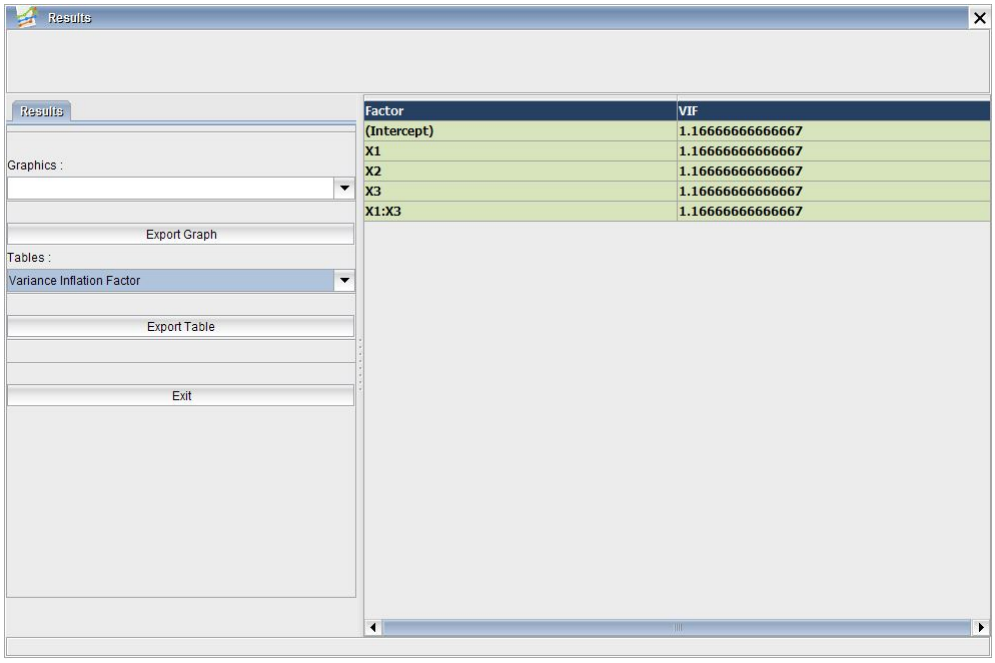


Figure 2.15: Variance Inflation Factor, Create Optimal Design

2.4 Full Longitudinal Design

A longitudinal design is an experimental design in which the researcher asses a particular group of people for some period of time. This kind of designs are special due to the correlation of the observations over time. Usually, this experiments are conducted to measures trends over a particular characteristic of interest of an individual or groups of individuals. The same people are studied in longitudinal studies so that the accuracy of the results are greatest. Longitudinal studies are also commonly used in the medical field because they help to measure medicine and discover certain diseases.

Summary of Full Longitudinal Design

- Full Longitudinal Designs are the easiest to construct. They are composed of all the combination of all the factors under study including Time. In addition, they are the easiest to analyze due to they are balanced designs.
- There are different ways to conduct the analysis of these designs. One is the subject-specific approach, where we are interested in modelling the response of each of the individuals in the study. Popular methods are *Linear Mixed Models*, *Generalized Linear Mixed Models*, etc. On the other hand, the population approach is used when we want to analyze the behaviour of whole groups. Popular methods are *Generalized Least Squares*, *Generalized Estimating Equations*.
- Incomplete Longitudinal Design are defined by using Cohorts. This methodology is important when we want to reduce the cost of our study. We will explain this in the next section.

Constructing a Full Longitudinal Design

As an example we have a longitudinal study of a weight-loss drug, administered at three dose levels, 0 (placebo), 1 and 2. The main goal is to test the alternative hypothesis that doses 1 and 2 bring about a weight loss trend different from the trend resulting from placebo. Two strains of mice were included in the study. Finally, the measures will be over 7 weeks of study.

In order to construct a Full Longitudinal Design for this study using LADES, we follow the route **Create** $->$ **Full Longitudinal Design**. In Figure 2.16 we appreciate the main screen of this function. We also see the two group characteristics under study with their corresponding levels. And we select the option **Randomize Design** in order to get our resulting design randomized over the individuals. We will see this in the Results section.

Moreover, the specifications for measurements over time have to be inputed in the **Time** section (button). See Figure 2.17. This full design will be the Case 1.

We will present the **Incomplete Design** characteristic as a the Case 2. Incomplete designs are useful when we want to reduce the total cost of our study. The term incomplete is used to represent the lack of some observations on specific time points. We can define more easily this concept using **Cohorts** schedule semantic showed on the right side of Figure **??**. The first row of this matrix corresponds to the first Cohort that represents that all individuals in this group will be evaluated at all seven occasions; we define the number of individuals in this group by inputting the desired number in the column **Size**. The second row represent the Cohort 2, where all the individuals of that group will be evaluated at occasions 1, 4 and 7. And Cohort 3 is composed of individuals that will be measured only at the first and last occasions. Varying the number of subjects in each cohort produces different incomplete designs. We can define different cohorts by selecting different groups of time points.

Function Outputs

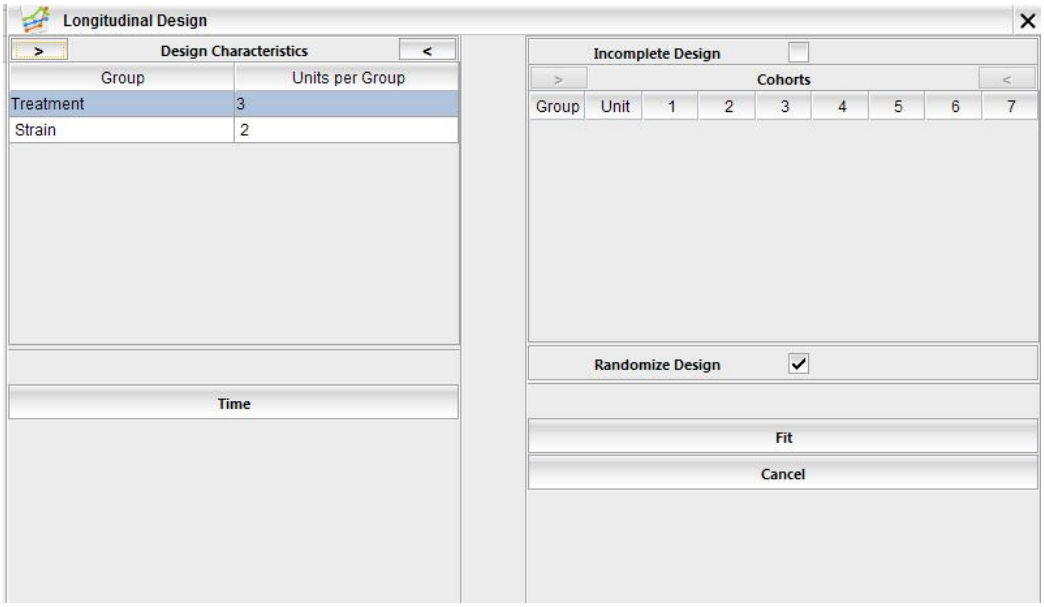


Figure 2.16: Main Screen, Full Longitudinal Design

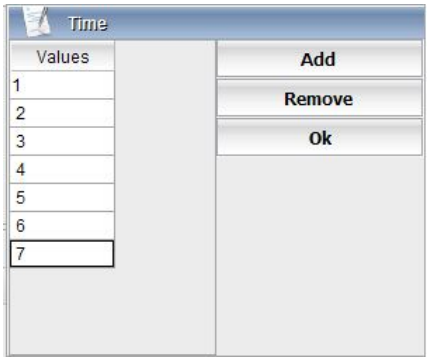


Figure 2.17: Time Measurements, Full Longitudinal Design

- **Longitudinal Design Case 1.** In this section of the Results output we see the full longitudinal design with the desired characteristics. A column showing the label for individuals (**Units**) was added. Since we choose to get an already randomiszd design, we appreciate that the design points are randomized within each level of **Time**. See Figure **??**.
- **Longitudinal Design Case 2.** Figure **??** presents the results of using the Incomplete Design characteristic. We can see that there are more observations for occasions 1 and 7, since the three cohorts included these occasions.

2.5 Unequal Longitudinal Design

The Unequal Longitudinal Designs (ULD) are based on constructing designs with unequal number of individuals per group. These unequal assignations can lead us to more power for the fixed effects test of hypothesis. This approach is proposed to reduce the total number of individuals used in a study, due to animal ethics, costs, etc. Their construction is very similar to the Longitudinal Optimal Powered-Cost Designs.

Summary of Longitudinal Optimal Powered-Cost Design

- A simple reparametrized linear mixed model is used to fit a model for each group under study.
- Linear Mixed Models provide an excellent framework to fit and evaluate the results of a longitudinal experiment. Moreover, the developed F_{Helms} test statistic let us evaluate fixed effects hypotheses and compute its power. Then, we can compare the designs according to its power.
- A simple cost function is defined in order to compare the constructed designs:

$$\text{cost} = N \times \text{MCS} \; + \; K \times \text{COS} \tag{2.2}$$

  Where N is the total number of subjects, K is the total number of observations per subjects, MCRS is the marginal cost of including (e.g. buying) an individual in the study, and COS is the cost of measure (maintain) and individual during K times. Other cost functions can be defined.
- Using this framework of power analysis, F_{Helms} statistic and cost function we can construct unequal longitudinal designs that reach the desired power for fixed effects hypothesis. This longitudinal designs are called *Unequal Longitudinal Design*. This designs look for a tradeoff between power and number of individuals, throughout cost function.
- It is important to have prior information in order to get more accurate results; Pilot studies, etc.

Constructing Unequal Longitudinal Designs

The sequence *Create* $->$ *Unequal Longitudinal Designs* gives the functionality for generating unequal designs. This function needs the parameter estimates, the power and cost of the original design that will be compared to the new designs, as inputs. As an example to show this methodology we have a mice study. Three treatments were administered to three groups of mice; Treatment A, Treatment B and a Placebo. Five measures over time were recorded and the response under study was the weight. The main goal of this example is to design the next experiment using the minimum number of mice and having a reasonable power for testing the difference among the slopes of three treatment.

The statistical model used by this function is a special case of a linear mixed model. This model has been reparametrized to make the comparisons of the three slopes more direct . The model for this example is:

$$y_{ij} = \begin{cases} (\beta_0 + b_{0i}) + (\beta_1 \, t_{ij} + b_{1i}) + \varepsilon_{ij}, & \text{Control} \\ (\beta_2 + b_{0i}) + (\beta_3 \, t_{ij} + b_{1i}) + \varepsilon_{ij}, & \text{T. B} \\ (\beta_4 + b_{0i}) + (\beta_5 \, t_{ij} + b_{1i}) + \varepsilon_{ij}, & \text{T. C} \end{cases} \tag{2.3}$$

where y_{ij} represents the j-ith measurement of the i-th eel. b_{0i} is the random factor that represents the difference of the initial weight to the mean; while b_{1i} is the random factor

representing the difference of each mouse to the average slope of each group. And ε_{ij} are independent of each other.

This model and the corresponding hypothesis testing for slopes was presented in Section 3.1.2 In our example we will use some of the results. We now present the function and how to fill the corresponding data.

First, Figure 2.18 shows the general parameter of the function such as **Number of Groups**, **Units per group**, and **Time**. We fill out the boxes with 3 groups, 10 mice, and 5 occasions (codified as -4, -2, 0, 2, 4), respectively. Minimum Units per Group refers to the minimum number of individuals the generated (new) designs must have; **Tolerance Difference Number for Individuals** stands for restricting possible (big) differences among the groups in the new generated designs.

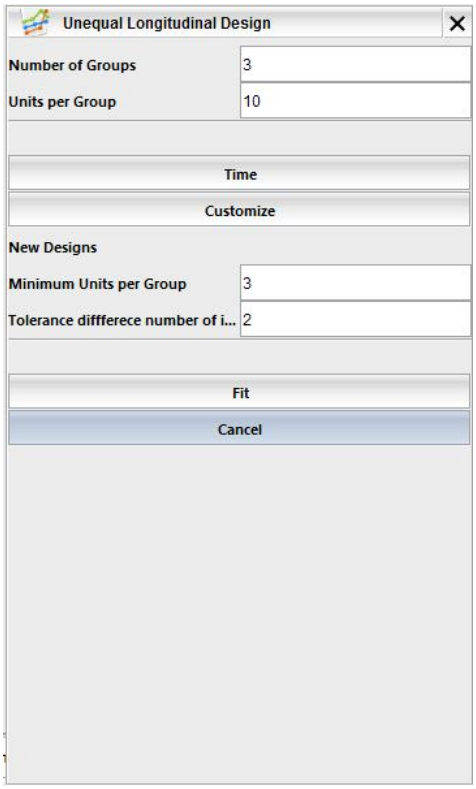


Figure 2.18: Main Screen, Unequal Longitudinal Designs

In the customize section we define the contrasts to use. On the left side of Figure 2.19 we define that we want to compare the group 1 (Control) against group 2 (T. A), and group 1 (T. A) against group 3 (T. B); using this definition we can complete a full comparison among the three groups. Now, we have to define the minimum difference to detect in these two comparisons. That is, the average slope of T. A. to the average slope of the control group, will be considered 'significative' if is at least -0.14 (we can also use 0.14). The same for the second contrast.

On the right side of Figure 2.19, we define if we want to discriminate the constructed designs based on cost. For this options we can use a **User's define Cost**; or, by leaving this option in zero, we discriminate against the value of the complete design (30 mice, 5 occasions). We also need to specify the cost of each unit (mouse), and the cost to get one measure; **Unit cost** and **Observation/Unit Cost**, respectively.

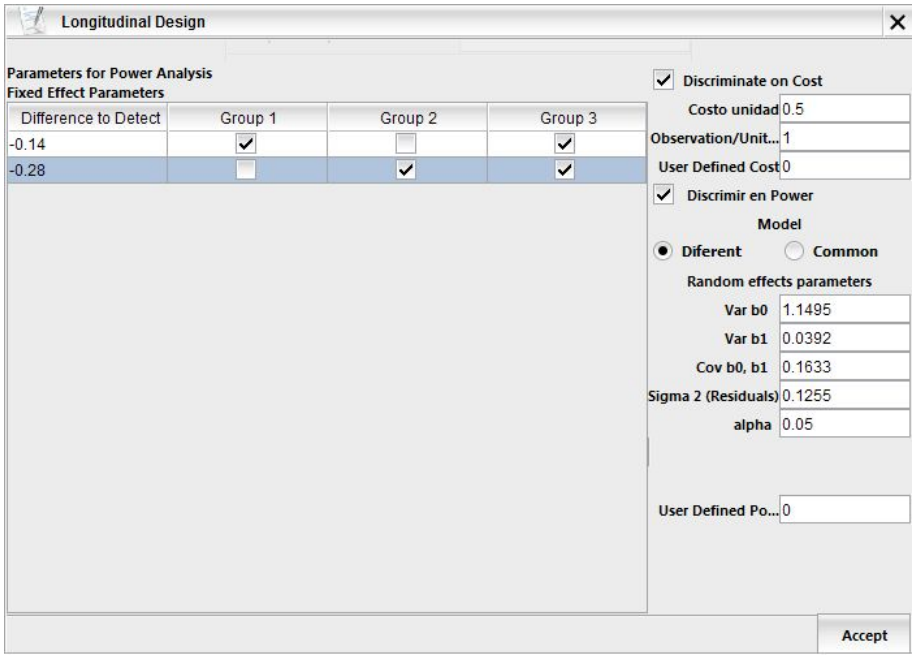


Figure 2.19: Customize Section, Unequal Longitudinal Designs

If we select the **Discriminate on Power** checkbox, we discriminate the generated designs based on a specific (desired) power. For this, we have to first define the type of model we are working with. **Different Intercept** means that each group model has an independent intercept (our case, See model 3.2); on the other hand, **Common** Intercept means that all group models will have an overall intercept representing the average of all mice's first observation.

Now, to define the random effects we are using in model (3.2), we simply fill out the boxes with the corresponding variances and covariances. If, for example, we fill the **Var b1** box with 0;, then the function will not include a random effect for slopes in the model. **Sigma 2**, stands for the variance of the residuals (σ_ε^2), and **alpha** represents type 1 error. When we want to discriminate the designs using the power of the complete design, we fill **User Defined Power** section with 0. If we are interested in some specific value, we just input it.

Finally, we just click on **Fit**, in the main screen.

Function Outputs

- ULD Design Graph:
- **Cost Efficient Designs by Cost**. Figure 2.20 depicts all designs with lower cost and more power than the original design (complete design).
- **Cost Efficient Designs by Units**. Figure 2.21 depicts all designs with minimum total number of animals used and more power than the original design (complete design).
- **Unequal Longitudinal Designs**. Table where the full information of the unequal longitudinal designs is shown. See Figure 2.22.
- **Reference Power**. Power of the complete design used as reference. See Figure2.23.
- **Reference Cost**. Cost of the complete design used as reference. See Figure 2.24.

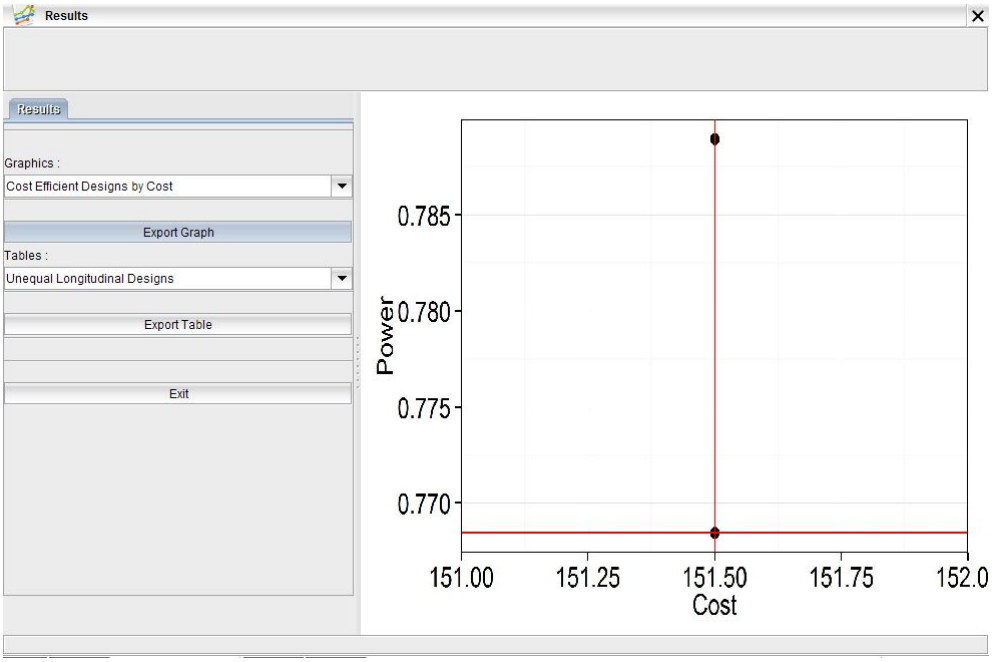


Figure 2.20: Cost Efficient Designs by Cost, design total cost on the X-axis and Total Power on the Y-axis

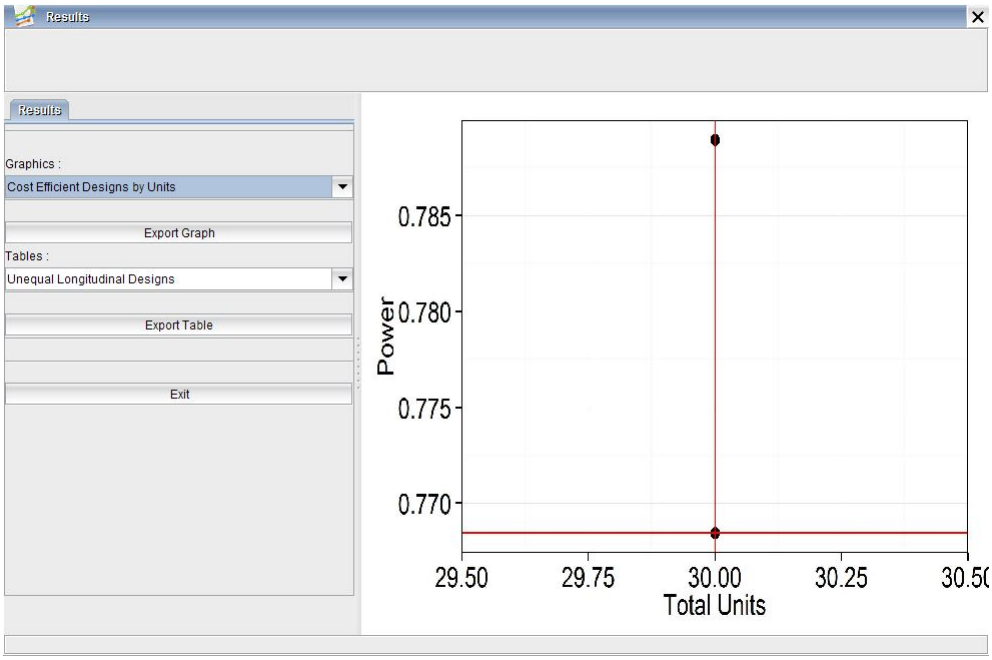


Figure 2.21: Cost Efficient Designs by Units, total number of units on the X-axis and Total Power on the Y-axis

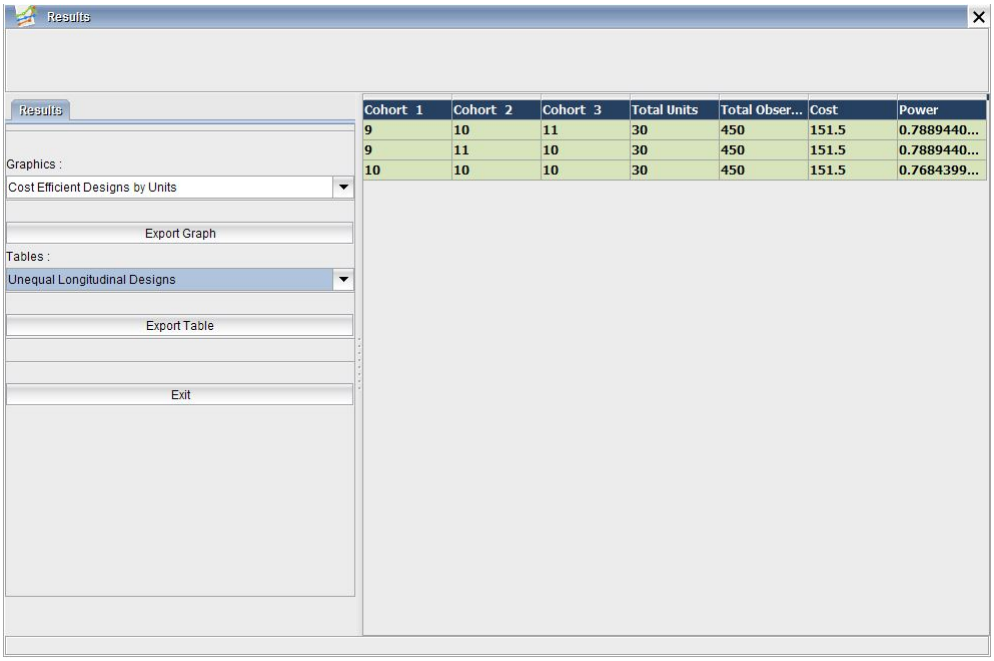

Results
Results
Graphics :
Cost Efficient Designs by Units
Export Graph
Tables :
Unequal Longitudinal Designs
Export Table
Exit
Cohort 1 Cohort 2 Cohort 3 Total Units Total Obser... Cost Power
9 10 11 30 450 151.5 0.7889440...
9 11 10 30 450 151.5 0.7889440...
10 10 10 30 450 151.5 0.7684399...

Figure 2.22: Unequal Longitudinal Designs, number of individuals per group, total cost and power

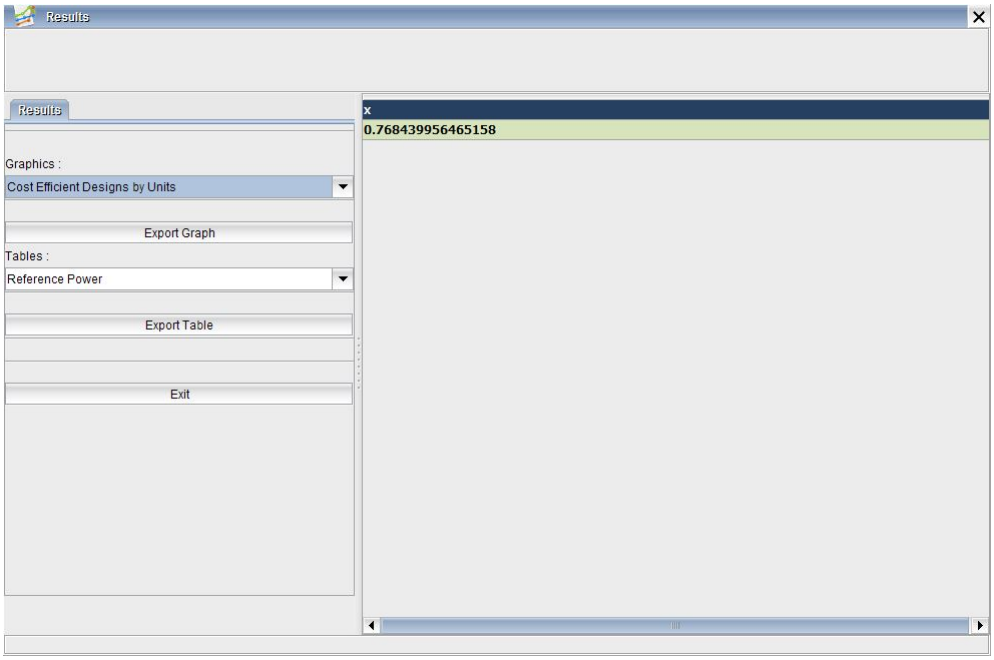


Figure 2.23: Reference Power, Unequal Longitudinal Designs

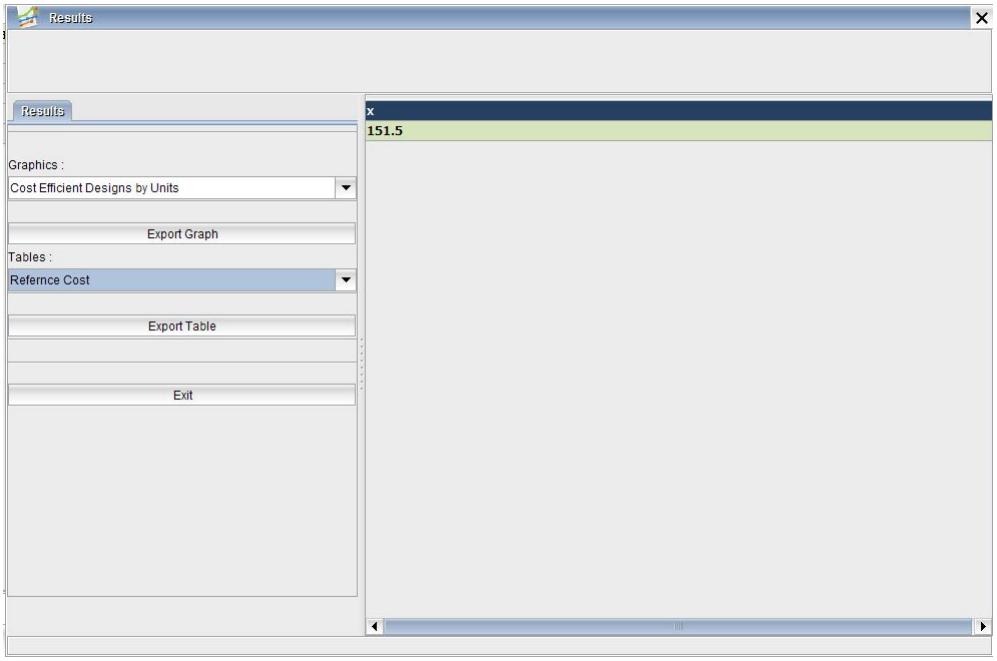


Figure 2.24: Reference Cost, Unequal Longitudinal Designs

2.6 Longitudinal Optimal Powered-Cost Design

Full longitudinal designs, in which each subject is evaluated at each measurement occasion, are the best option for conducting longitudinal studies but are often very expensive and motivate a search for more efficient designs. Intentionally incomplete designs represent less cost since it divides the population under study in groups that are measured only on specific occasions, called *Cohorts*. This type of designs have the potential to be more efficient than complete designs.

Summary of Longitudinal Optimal Powered-Cost Design

- Linear Mixed Models provide appropriate analysis tools for this type of designs.
- Fixed effect hypotheses can be tested via the F_{Helms} test statistics. An accurate approximation of the statistic's small sample non-central distribution makes power computation feasible. Thus, we can compare the longitudinal designs according to the power of fixed effects test of interest under the F_{Helms} statistic.
- A simple cost function is defined in order to compare the constructed design.

$$\text{cost} = N \times \text{MCS} \; + \; K \times \text{COS} \tag{2.4}$$

  Where N is the total number of subjects, K is the total number of observations per subjects, MCRS is the marginal cost of including (e.g. buying) an individual in the study, and COS is the cost of measure (maintain) and individual during K times. Other cost functions can be defined.
- Using this framework of power analysis, F_{Helms} statistic and cost function we can construct incomplete longitudinal designs that reach the desire power for fixed effects hypothesis. We call this longitudinal designs, *Longitudinal Optimal Powered-Cost Design*. This designs look for a tradeoff between power and cost.
- We need prior information in order to get more accurate results.

Constructing Longitudinal Optimal Powered-Cost Design

As an example to show this methodology we have a mice study. Three treatments were administered to three groups of mice; Treatment A, Treatment B and a Placebo. Five measures over time were recorded and the response under study was the weight. The main goal of this example is to design the next experiment using the minimum number of mice and having a reasonable power for testing the difference among the slopes of three treatment. We will explain the example and some important characteristics of this function. This function can be accessed by following: **Create** $->$ **Longitudinal Optimal Powered-Cost Design**.

The statistical model used by this function is a special case of a linear mixed model. This model has been reparametrized to make the comparisons of the three slopes more direct . The model for this example is:

$$y_{ij} = \begin{cases} (\beta_0 + b_{0i}) + (\beta_1\, t_{ij} + b_{1i}) + \varepsilon_{ij}, & \text{Control} \\ (\beta_2 + b_{0i}) + (\beta_3\, t_{ij} + b_{1i}) + \varepsilon_{ij}, & \text{T. B} \\ (\beta_4 + b_{0i}) + (\beta_5\, t_{ij} + b_{1i}) + \varepsilon_{ij}, & \text{T. C} \end{cases} \tag{2.5}$$

where y_{ij} represents the j-ith measurement of the i-th eel. b_{0i} is the random factor that represents the difference of the initial weight to the mean; while b_{1i} is the random factor representing the difference of each mouse to the average slope of each group. And εij are independent of each other.

This model and the corresponding hypothesis testing for slopes was presented in Section 3.1.2. In our example we will use some of the results. We now present the LADES function and how to fill the corresponding data.

First, Figure 2.25 shows the general parameter of the function such as **Number of Groups**, **Units per group**, and **Time**. We fill out the boxes with 3 groups, 10 mice, and 5 occasions, respectively. We leave the **Customize** section for a moment, and we will focus on the **Cohorts** section. In the Cohorts section we can define up to a maximum of 3 cohorts. Here, we can add cohorts, select the occasions that conform each cohort, and we have to define separately the number of observations regarding cohorts. For example, the second row in Figure **??** shows that the second Cohort is composed of the first, the last and the middle occasions, and the total number of occasions was 3. Researchers decided to include three cohorts to reduce the number of subject occasions and cost.

In the customize section we define the contrasts to use. On the left side of Figure **??** we define that we want to compare the group 1 (Control) against group 2 (T. A), and group 2 (T. A) against group 3 (T. B); using this definition we can complete a full comparison among the three groups. Now, we have to define the minimum difference to detect in these two comparisons. That is, the average slope of T. A. to the average slope of the control group, will be considered 'significative' if is at least -0.14 (we can also use 0.14). The same for the second contrast.

On the right side of Figure 2.26, we define if we want to discriminate the constructed designs based on cost. For this options we can use a **User's define Cost**; or, by leaving this option in zero, we discriminate against the value of the complete design (30 mice, 5 occasions). We also need to specify the cost of each unit (mouse), and the cost to get one measure; **Unit cost** and **Observation/Unit Cost**, respectively.

If we select the **Discriminate on Power** checkbox, we discriminate the generated designs based on a specific (desired) power. For this, we have to first define the type of model we are working with. **Different Intercept** means that each group model has an independent intercept (our case, See model 3.2); on the other hand, **Common** Intercept means that all group models will have an overall intercept representing the average of all mice's first observation.

Now, to define the random effects we are using in model (3.2), we simply fill out the boxes with the corresponding variances and covariances. If, for example, we fill the **Var b1** box with

Figure 2.25: Main Screen, Longitudinal Optimal Powered-Cost Design

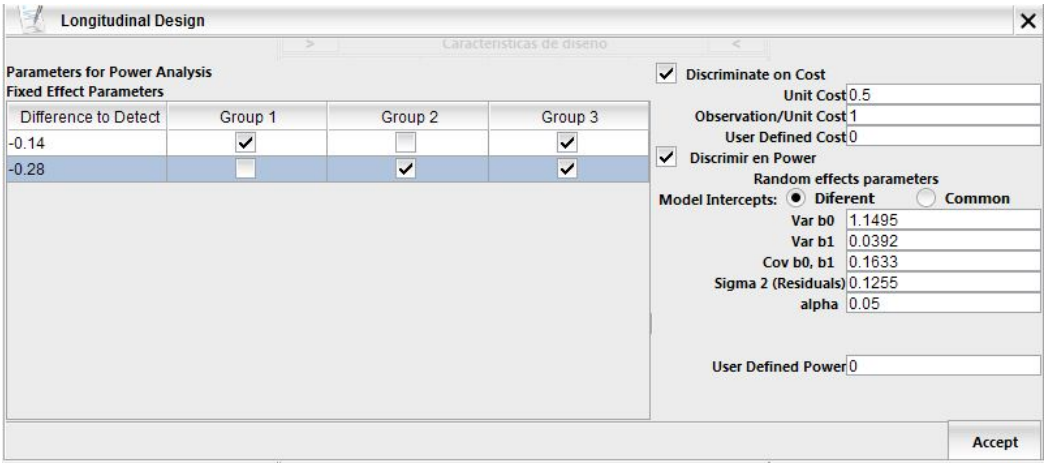


Figure 2.26: Customize Section, Longitudinal Optimal Powered-Cost Design

0;, then the function will not include a random effect for slopes in the model. **Sigma 2**, stands for the variance of the residuals, and **alpha** represents type 1 error. When we want to discriminate the designs using the power of the complete design, we fill **User Defined Power** section with 0. If we are interested in some specific value, we just input it.

Finally, we just click on **Fit**, in the main screen.

Function Outputs

- **Reference Cost.** This section shows the maximum cost (according to eq. 3.1) used to discriminate the constructed designs. See Figure 2.27.

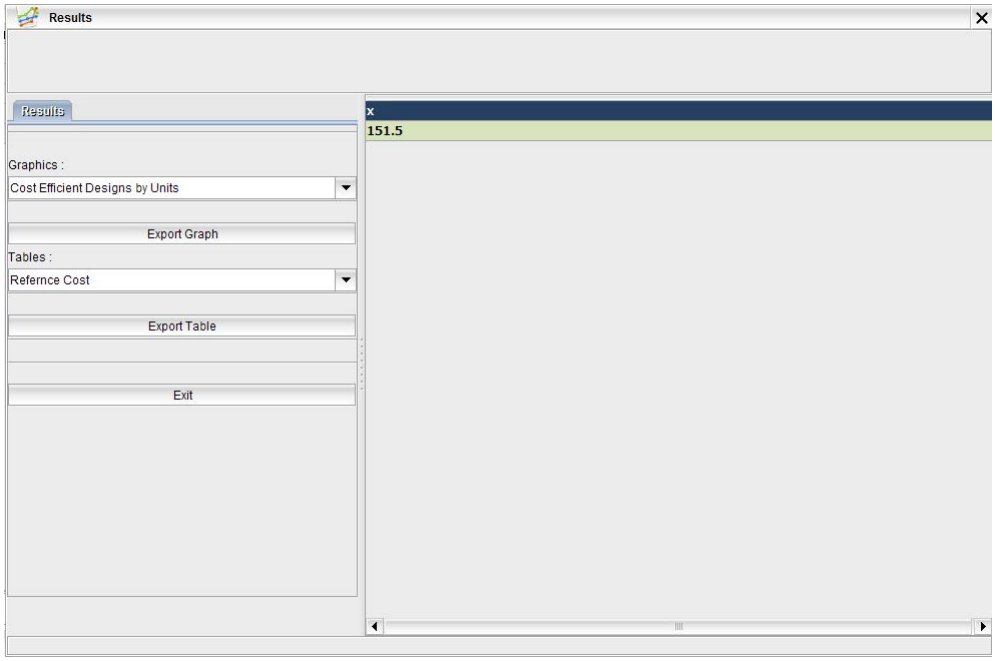


Figure 2.27: Reference Cost, Optimal Cost-Efficient Longitudinal Designs

- **Reference Power.** This section shows the minimum power used to discriminate the constructed designs. See Figure 2.28.
- **Longitudinal Optimal Design.** Shows all the constructed designs with the desired power and cost. See Figure 2.29.
- **Power vs Cost Graph.** This graphs gives us a perspective of all the designs having the desired properties. It shows cost on X axis, and Power on y axis. See Figure 2.30
- **Power vs Total Units.** Presents Total Number of Units on X axis, and Power on y axis. See Figure 2.31.

We encourage the reader to check the paper [**Helms:1992aa**], for more information about the methodology.

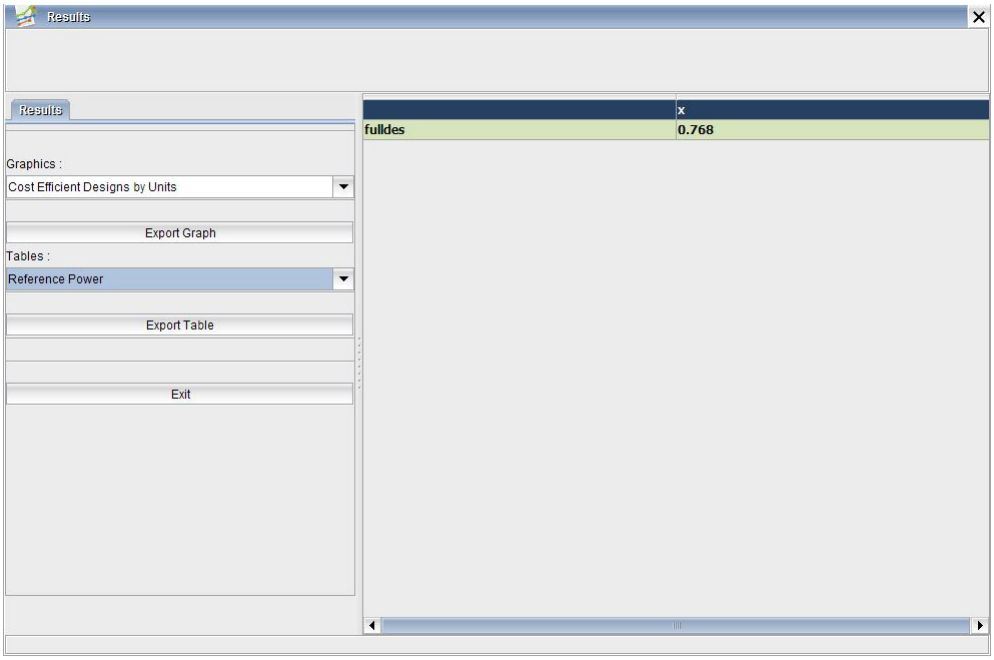


Figure 2.28: Reference Power, Optimal Cost-Efficient Longitudinal Designs

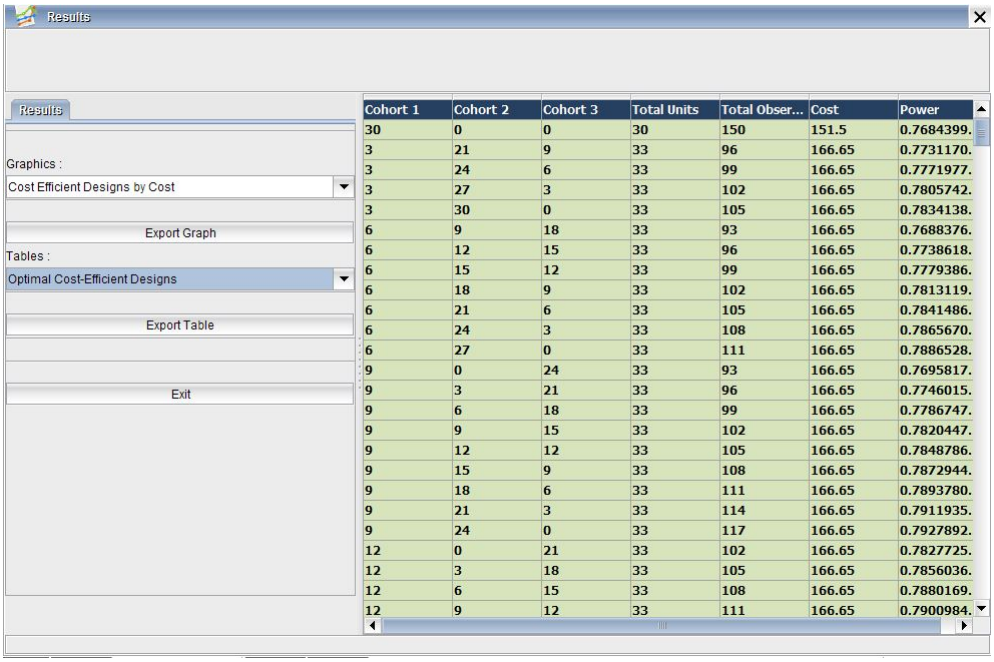


Figure 2.29: Optimal Cost-Efficient Longitudinal Designs, number of individuals per group, total cost and power

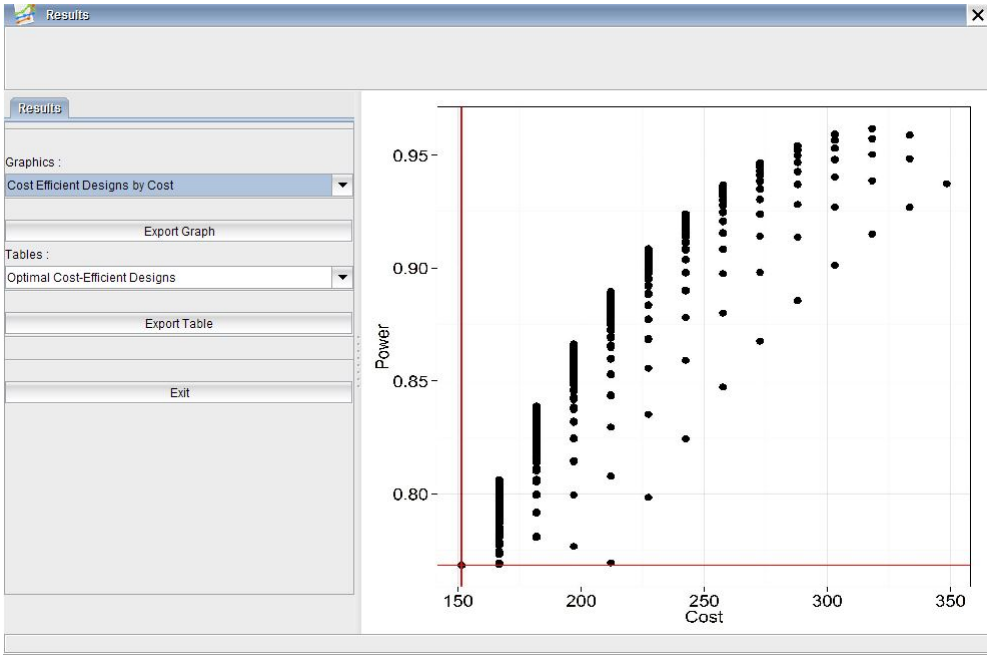


Figure 2.30: Cost Efficient Designs by Cost, design total cost on the X-axis and Total Power on the Y-axis

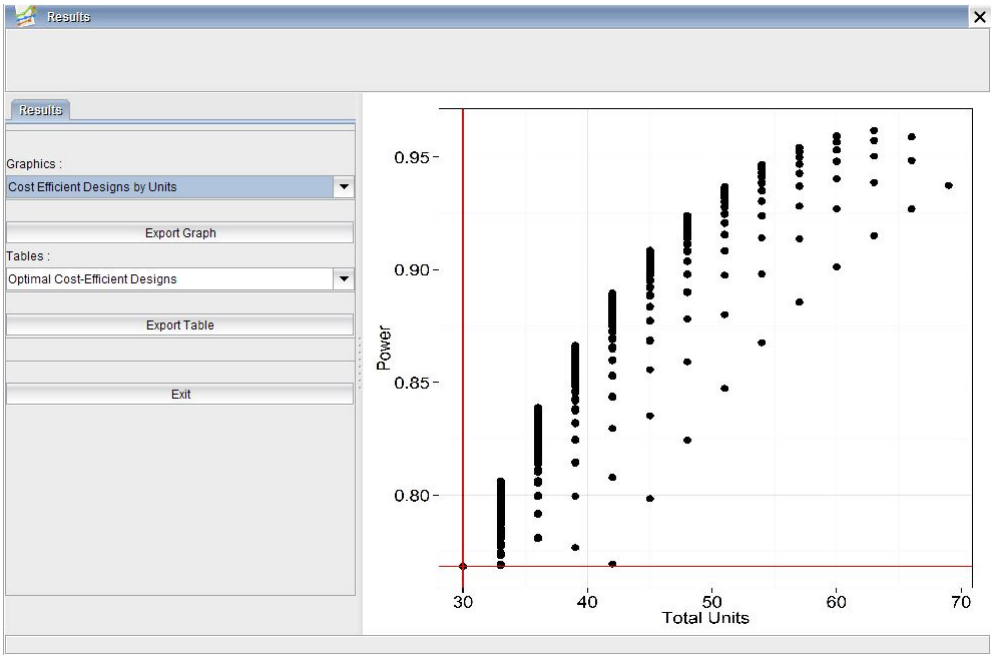
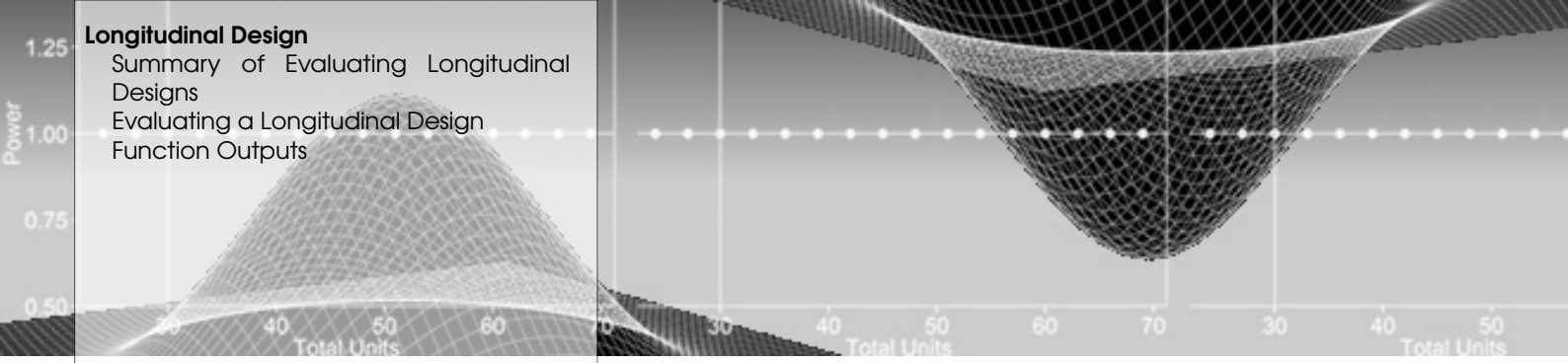


Figure 2.31: Cost Efficient Designs by Units, total number of units on the X-axis and Total Power on the Y-axis

Longitudinal Design
Summary of Evaluating Longitudinal Designs
Evaluating a Longitudinal Design
Function Outputs



3 — Evaluate Module

3.1 Longitudinal Design

Once we have detecting a significant difference among treatments, we have to evaluate its power. Power is the probability of detecting a significative difference when it really exist. Power for the most common test such as t-test or anova have been developed, but power for longitudinal parameters is still under study. This function uses a method to evaluate the power of a test for the average slopes, basing on a linear mixed model framework.

3.1.1 Summary of Evaluating Longitudinal Designs

- Linear Mixed Models provide appropriate analysis tools for this type of designs.
- Fixed effect hypotheses can be tested via the F_{Helms} test statistics. An accurate approximation of the statistic's small sample non-central distribution makes power computation feasible. Thus, we can compare the longitudinal designs by the power of the fixed effects test of interest under the F_{Helms} statistic.
- A simple cost function is defined in order to compare the constructed design.

$$\text{cost} = N \times \text{MCS} \; + \; K \times \text{COS} \tag{3.1}$$

  Where N is the total number of subjects, K is the total number of observations per subjects, MCRS is the marginal cost of including (e.g. buying) an individual in the study, and COS is the cost of measure (maintain) and individual during K times. Other cost functions can be defined.
- We need prior information in order to get more accurate results, such as random effects variances, and residual variances.
- Under this framework, we can also evaluate Intentionally Longitudinal Design.

3.1.2 Evaluating a Longitudinal Design

As an example to show this methodology we have a mice study worked in [**Helms1992**]. Three treatments were administered to three groups of mice; Treatment A, Treatment B and a Placebo. Five measures over time were recorded and the response under study was the weight of mice. The main goal of this example is to test the hypothesis for average slopes and evaluate its power.

We will explain the example and some important characteristics of this function. This function can be accessed by following: **Evaluate** $->$ **Longitudinal Design**.

The statistical model used by this function is a special case of a linear mixed model. This model has been reparametrized to make the comparisons of the three slopes more direct . The model for this example is:

$$y_{ij} = \begin{cases} (\beta_0 + b_{0i}) + (\beta_1\, t_{ij} + b_{1i}) + \varepsilon_{ij}, & \text{Control} \\ (\beta_2 + b_{0i}) + (\beta_3\, t_{ij} + b_{1i}) + \varepsilon_{ij}, & \text{T. B} \\ (\beta_4 + b_{0i}) + (\beta_5\, t_{ij} + b_{1i}) + \varepsilon_{ij}, & \text{T. C} \end{cases} \tag{3.2}$$

where y_{ij} represents the j-ith measurement of the i-th eel. β_0, β_2, and β_4 represent the average slope for each group; and, β_1, β_3, and β_5 represent the average slope for their corresponding group. b_{0i} is the random factor that represents the difference of the initial weight to the mean; while b_{1i} is the random factor representing the difference of each mouse to the average slope of each group. And ε_{ij} are independent of each other. The levels for time were codified as $-4,-2,0,2,4$, so the model is evaluating the decrease in weight around the third week.

After fitting the linear mixed model the estimates were: β_0=-1.3875, β_1= -0.0350, β_2=-2.3875, β_3= -0.17625, β_4=-3.3750, β_5=-0.3175; σ_{b_0}=1.1495, σ_{b_1}=0.0392, σ_{b_0,b_1}=0.163375; and σ_ε=0.1255.

The hypothesis test we want to asses is:

$$H_0 : \beta_1 = \beta_3 = \beta_5 \;\text{ vs }\; H_1 : \text{at least one different}, \tag{3.3}$$

representing that we want to test the average slopes of groups. A level of $\alpha = 0.05$ will be used. Under this framework, we will use the the F approximation by Helms to test the hypothesis and compute its power. More information about the theory and methods can be found in Helms ([**Helms1992**]). Now we will present how to fill out this data in the corresponding boxes.

First, Figure 3.1 shows the general parameter of the function such as **Design Characteristics** (3 groups, 10 mice), where we can add up to 3 groups to evaluate the hypothesis using F_Helms. Below this section we found the **Time** section where we define the time used (5 occasions); and, the **Cost Evaluation** section if we want to calculate the cost of our design.

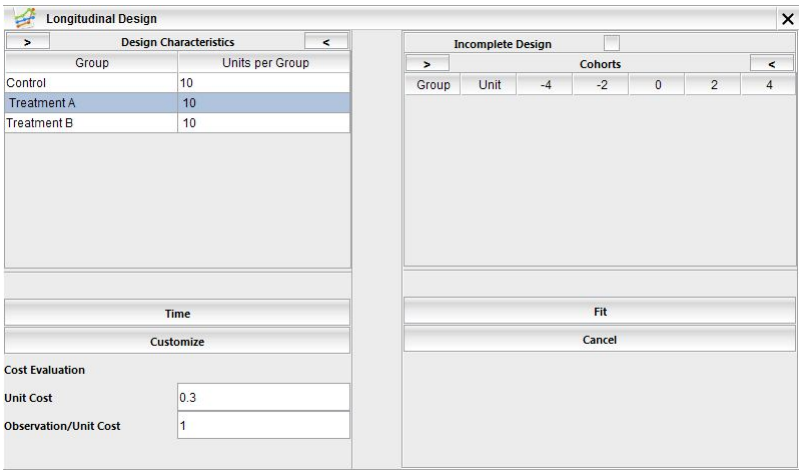


Figure 3.1: Main Screen, Evaluate Longitudinal Design

We can define an Incomplete Design, also. We just need to select the Group and add up to 3 *Cohorts* and define the number of individuals for each. We can see an example in Figure 3.2. For instance, the second row in Figure 3.2 shows that the second Cohort is composed of the first, the last and the middle occasions, and the total number of occasions was 3. Researchers decided to include three cohorts to reduce the number of subject occasions and cost.

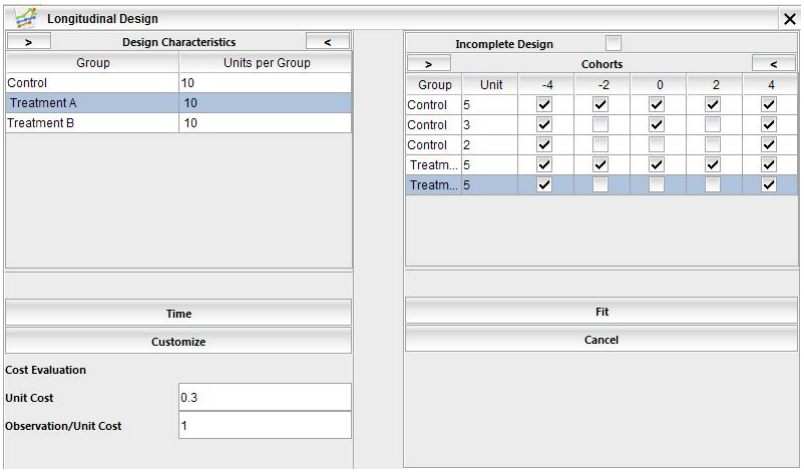


Figure 3.2: Main Screen using Incomplete Longitudinal Designs, Evaluate Longitudinal Design

Now, in the customize section we define the contrasts to use and the parameter estimates (Fig 3.3). In the **Fixed Effects Parameters**, we input the corresponding parameters for fixed parameters (intercepts and slopes). We can see that the first column of this matrix contains a row called **Beta Global**; here we can input the parameter estimation if we were working with a linear mixed effect model with a global intercept for all groups. On the left side of this screen we have to input the parameters for variances of random effects and residuals. If, for example, we fill the **Var b1** box with 0;, then the function will not include a random effect for slopes in the model. **Sigma 2**, stands for the variance of the residuals, and **alpha** represents type 1 error. When we want to discriminate the designs using the power of the complete design, we fill **User Defined Power** section with 0. If we are interested in some specific value, we just input it.

In the **Slopes Comparison** section we define the contrast we will use. We define that we want to compare the group 1 (Control) with group 2 (T. A), and group 1 (Control) with the group 3 (T. B); using this definition we can complete a full comparison among the three groups. With these characteristics, the function will define the minimum difference to detect in these two comparisons. The same holds for the **Intercept Comparison** section, if we leave this section in zero, then we will not compute any test.

Finally, we just click on **Fit**, in the main screen.

3.1.3 Function Outputs

- **Design Cost.** This section shows the total cost of the design (according to eq. 3.1).
- **Power Slopes.** This section shows the power for the slopes test of hypothesis.
- **Power Intercept.** This section shows the power for the intercept test of hypothesis.
- F_{Helms} **and Test.** Shows the results of the test of hypothesis (3.3).

We encourage the reader to check the paper [**Helms1992**], for more information about the methodology.

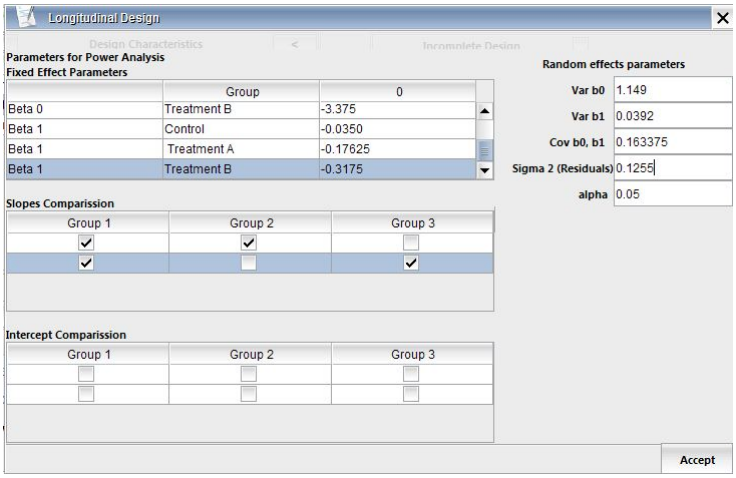
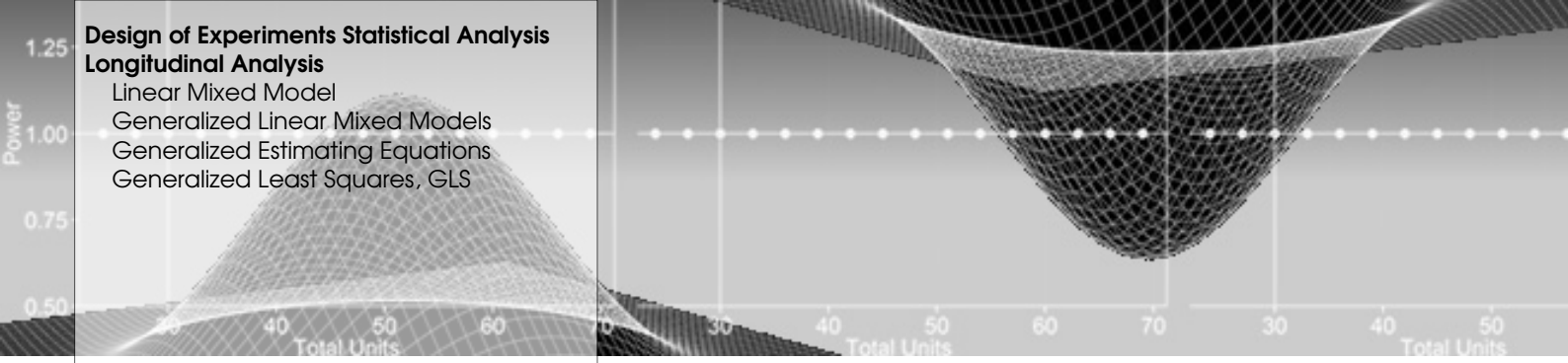


Figure 3.3: Customize Section, Evaluate Longitudinal Design



4 — Analyze Module

4.1 Design of Experiments Statistical Analysis

After running the experimental design, we have to analyze the results in an adequate statistical manner. We want to test the parameter effects and to construct a model that better represents our data. The analysis we use is similar to multiple linear regression.

Summary Design of Experiments Statistical Analysis

- The main effect of a factor is the resulting difference of subtracting the average of the observations in its lower level minus the average of the observations in its higher level. The main effects graph shows the average of all observations according to each level in each factor, and then connect them with a line. The estimated effect of the interaction of two factors, A and B, is calculated as the resulting difference of subtracting: the average of the observations of factor B at its highest level with the factor B at its low level, both given the highest level A; minus, the difference of factor B at its highest with factor B in low level, both at the lower level of A. The effects of interactions can be seen through an interaction plot. This graph shows the average of the observations in each combinations of two factors A and B [(-, -), (+, -), (-, +), (+, +)] and these averages are joined by a line.
- The model matrix helps us to compute main effects and interactions effects. These effects can be estimated using **linear regression** (See Section 5.1); we only need the model matrix and response vector (See A.1). By least squares the coefficients are estimated for each of the dependent variables (factors) of the regression model. These coefficients multiplied by 2 are the resulting effects for each factor (main and interactions).
- We can use the analysis of variance to assess which effects are statistically significant. Since the ANOVA is based on the F-test and the degrees of freedom of the residuals, we would prefer a replicated design in order to have efficient results. When such a replicated design is not possible we can use the Half-Normal plot. The Half-Normal plot is a graph where the absolute value of the estimates of the effects and its normal cumulative probabilities are plotted. We may also use the t-test for regression coefficients and decide which coefficients are statistically significant in the model; but this test is equivalent to ANOVA.
- There are three principles in the design of experiments that we must take into account at the time of our analysis. **Principle of Hierarchy**: (i) is more likely that the effects of low-

order (e.g. main) are more important than higher order effects (e.g. interactions, quadratic effects, etc.), (ii) the effects of the same order are equally important. **Sparcity Principle**: the number important factors in a factorial experiment is small. **Principle Heredity**: For significant interaction, at least one of the factors involved should be significant.

- To verify if our model is correct, i.e. the factors involved in our model are really significant for our response; we have to check some assumptions. It is assumed that the residuals of our model: 1) follow a Normal distribution; 2) have constant variance; and 3) are independent. This means, in other words, the residuals have no more information.

Implementation of an Analysis of an Experimental Design

As an example we will analyze the results of an study about two different strains of mice, BALB/c y C57BL. To each of the strains, a treatment and a placebo were administered. the response under study was the number of red blood cells (RBC). Researchers are interested in assessing the difference between the number of red cells between strains, and a possible difference between the strains under the new treatment (strain-treatment interaction). The design that was used for this experiment is shown in Table 4.1.

Cepa	**Tratamiento**	**RBC**
BALB/c	Placebo	10.1
BALB/c	Placebo	9.73
C57BL	Placebo	9.2
C57BL	Placebo	9.14
BALB/c	Placebo	10.08
C57BL	Placebo	9.6
BALB/c	Treatment	8.68
BALB/c	Treatment	8.45
C57BL	Treatment	8.18
BALB/c	Treatment	8.95
BALB/c	Treatment	8.89
C57BL	Treatment	8.82
BALB/c	Placebo	10.09
C57BL	Treatment	8.1
C57BL	Treatment	8.24
C57BL	Placebo	9.56

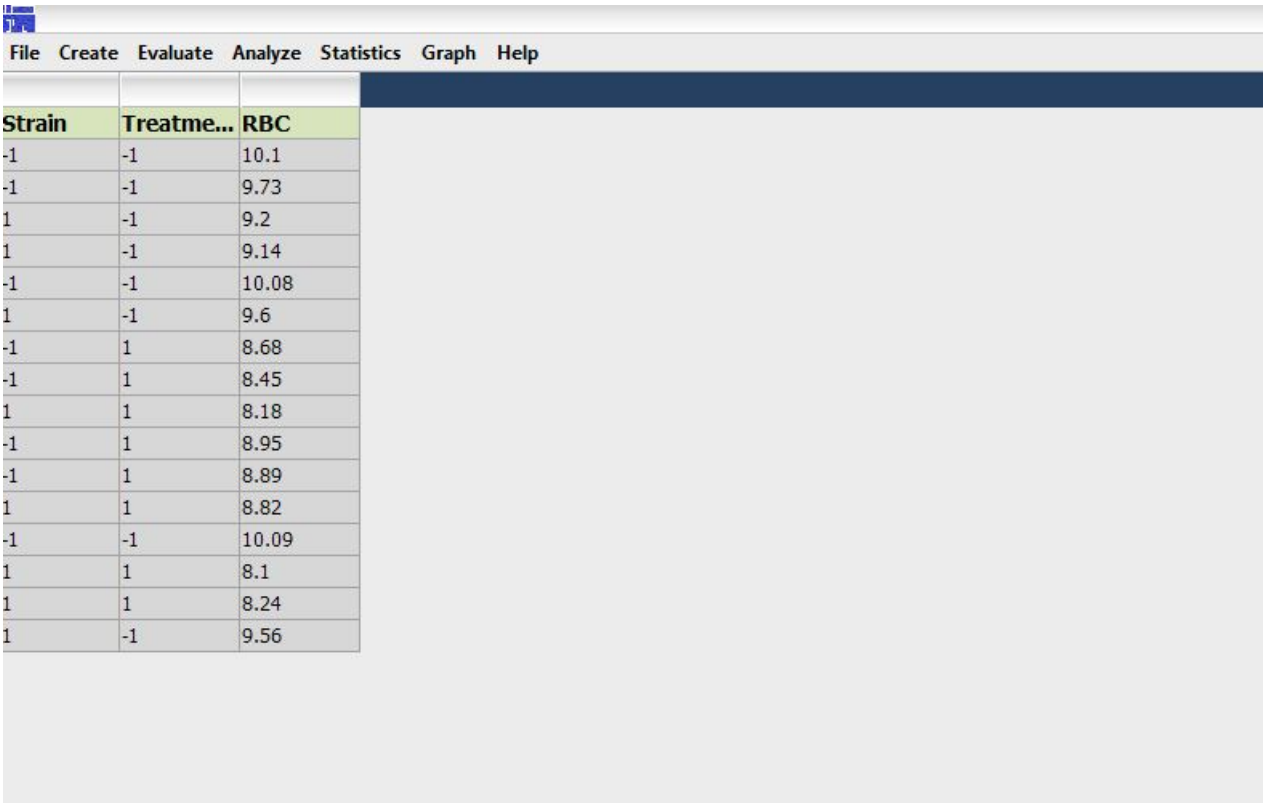
Table 4.1: RBC Experimental Data

When analyzing an experimental design, it is important to codify the variables to construct an ortogonal design matrix. This can lead us to a more accurate statistical analysis. The encoded levels for each of our variables are shown in Table .

Strain	**Codification**
BALB/c	-1
C57BL	1
Treatment	**Codification**
Placebo	-1
Treatment	1

Table 4.2: Codifications for each Factor in the RBC Experiment

Data imported to the software is shown in Figure 4.1. Here we can see that it is a 2^2 full factorial design replicated 3 times. Now, to access the **Design of Experiments Analysis** function, we follow **Analyze** $->$ **DOE**. The resulting screen is shown in Figure 4.2.



Strain	Treatme...	RBC
-1	-1	10.1
-1	-1	9.73
1	-1	9.2
1	-1	9.14
-1	-1	10.08
1	-1	9.6
-1	1	8.68
-1	1	8.45
1	1	8.18
-1	1	8.95
-1	1	8.89
1	1	8.82
-1	-1	10.09
1	1	8.1
1	1	8.24
1	-1	9.56

Figure 4.1: RBC Data. File **RBC.csv**

In Figure 4.2 the proposed model is shown. This model has the main factors of **strain** and **treatment**, and their interaction.

Function Outputs

- **Hypothesis Testing for Parameters**. The hypothesis test were assessed using the t-statistic. The tesi evaluates if the coefficient is different from zero. Only the main factors **strain** and **treatment** have a resulting α below 0.05. Therefore, we can say these factors have a significant impact on the response, number of red blood cells; i.e. RBC is different for strains, as well as treatments (see Figure 4.3).
- **Model Matrix**. In this section we can see the resulting model matrix (see A.1). This matrix was used to adjust the multiple linear regression model to our design (Figure 4.4). As we can see, this matrix is orthogonal; so, we can ensure that our estimates for the coefficients are the most accurate.
- **Coefficients**. Displays the estimates of the coefficients in our regression model (Figure 4.5). This allows us to define an equation that represents our data, and then make predictions on new data. The model in this case would be (using encodings in Table 4.1).

$$RBC = 9.113 - 0.258\,Strain - 0.574\,Treatment + 0.054\,Treatment : Strain \quad (4.1)$$

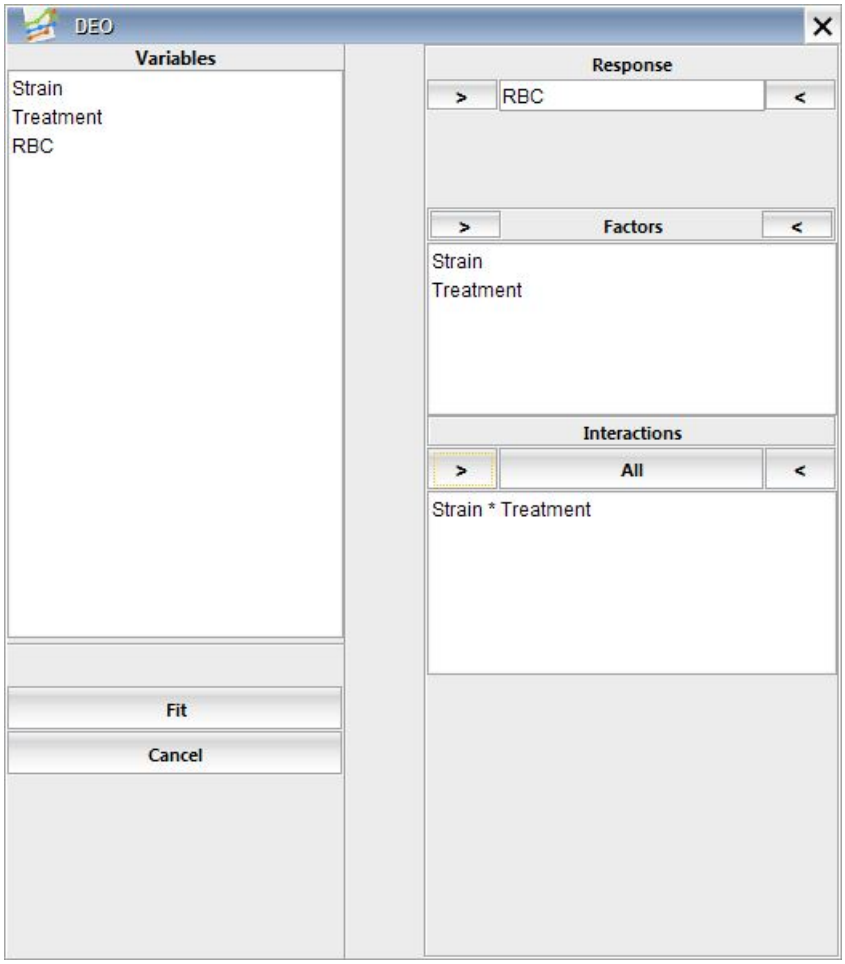


Figure 4.2: Analyze DOE Screen Function

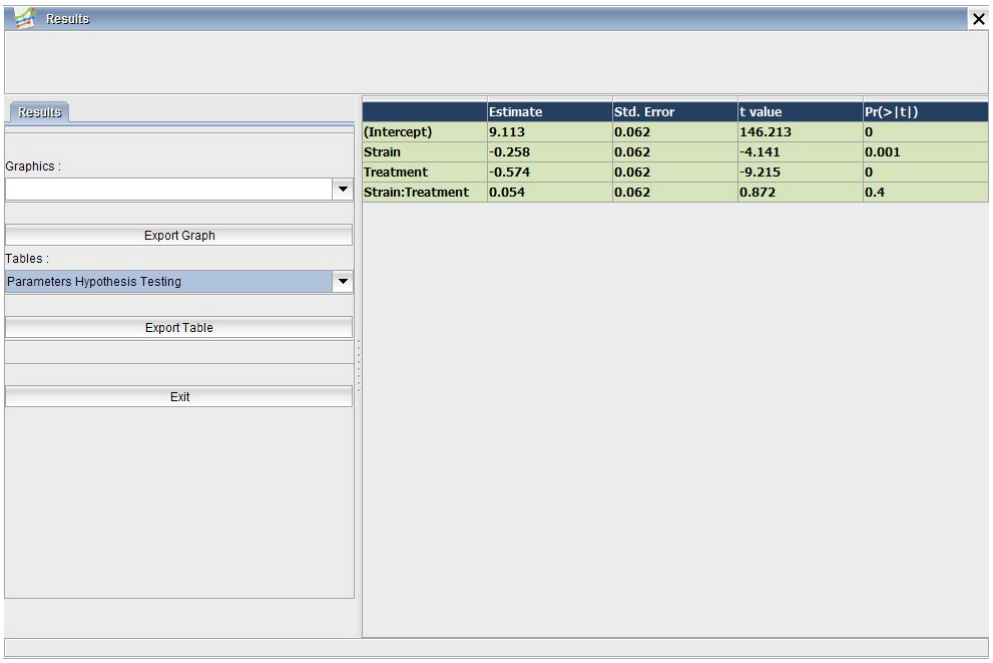


Figure 4.3: Hypothesis Testing for Parameters, DOE Analysis

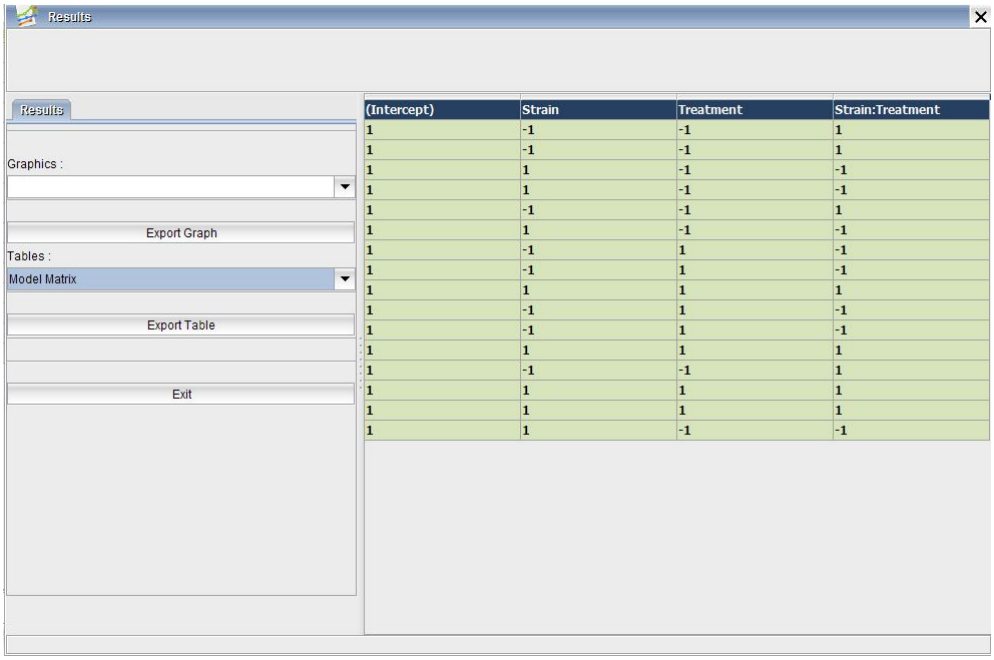


Figure 4.4: Model Matrix, DOE Analysis

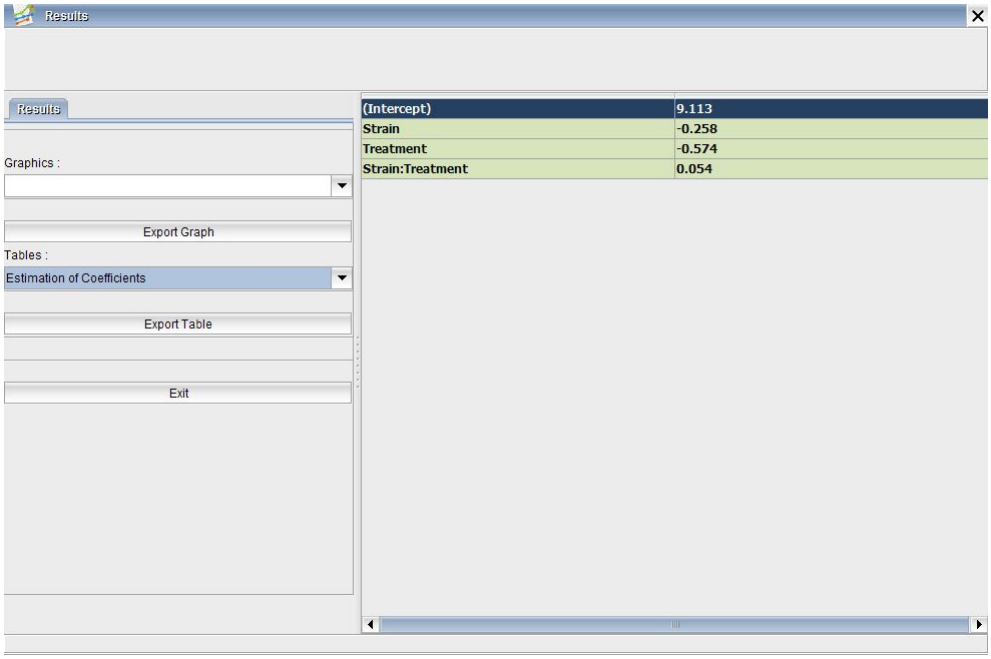
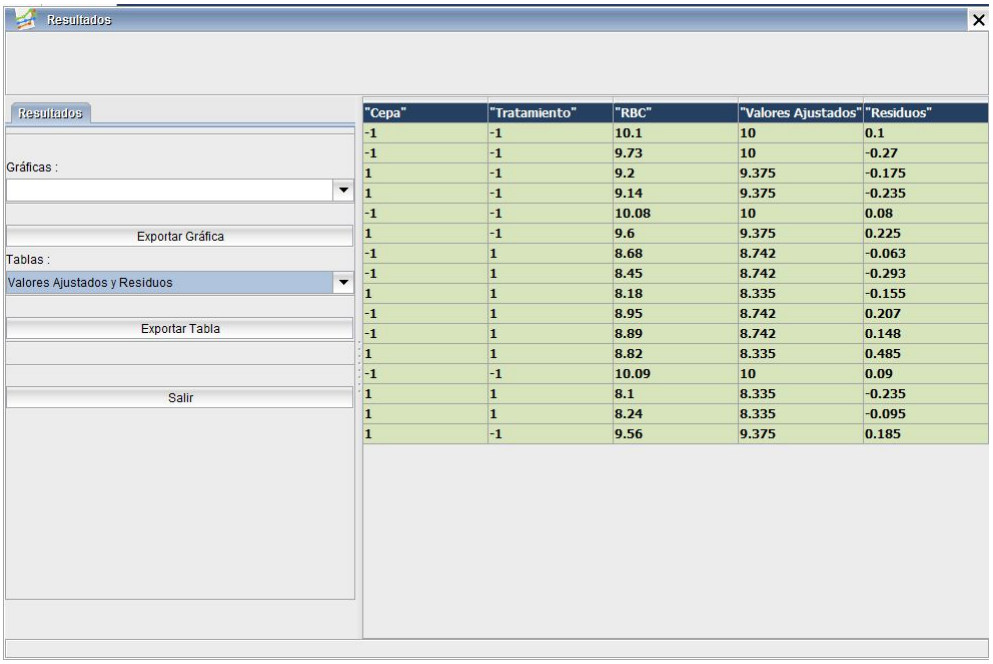


Figure 4.5: Coefficients, DOE Analysis

- **Adjusted Values and Residuals**. Displays a table containing regressors, predicted values and residuals (actual minus predicted) (4.1). See Figure 4.6.
- **R-Square, Adjusted R-Square, and Standard Error**. The multiple R-square indicates the proportion of variation in the dependent variable explained by the set of independent (predictor) variables. This statistics is used to check the goodness of fit of model (4.1). For our case, it showed that multiple R-square = 0.8955. The closer 1, the better the goodness of fit. See Figure 4.7.
- **Analysis of Variance**. Shows the sum of squares for each of the factors as well as its F-statistics to assess their significance (Figure 4.8). It clearly shows the significance of **Treatment** and **Strain** factors (since the resulting p-value is a lower than $\alpha = 0.05$). These data ais consistent with the results shown in Figure . However, ANOVA is considered as the most reliable test to determine factor significance.
- **Quantile-Quantile Normal Plot**. Graph where residuals empirical quantiles and standard normal quantiles are plotted (Figure 4.9). It helps us to validate the assumption of **normality, i)**. We aim to have all points very close to the midline. This chart helps us to detect possible outliers that occurred in our study. For our data, we can see that most falls around the midline. We can detect, however, an atypical point in the lower left corner of the chart.
- **Residuals Independence Graph**. This graph shows the times (X axis) in which each observation was collected and the residuals of each observation (Y axis). It helps us to validate the assumption of **independence, iii)**. This graph shows no trend, up or down, i.e. is around zero. If you have an upward trend, for example, this would provide evidence that the current observation depends on the former. For our data, it appears that there is no explicit trend. See Figure 4.10.
- **Residuals VS Fitted Plot**. Adjusted Model Values are plotted on the X axis, and the corresponding residuals on the Y axis. It helps us to validate the constant **variance assumption ii)**. It is desired the points on the graph form an horizontal band; i.e. the



"Cepa"	"Tratamiento"	"RBC"	"Valores Ajustados"	"Residuos"
-1	-1	10.1	10	0.1
-1	-1	9.73	10	-0.27
1	-1	9.2	9.375	-0.175
1	-1	9.14	9.375	-0.235
-1	-1	10.08	10	0.08
1	-1	9.6	9.375	0.225
-1	1	8.68	8.742	-0.063
-1	1	8.45	8.742	-0.293
1	1	8.18	8.335	-0.155
-1	1	8.95	8.742	0.207
-1	1	8.89	8.742	0.148
1	1	8.82	8.335	0.485
-1	-1	10.09	10	0.09
1	1	8.1	8.335	-0.235
1	1	8.24	8.335	-0.095
1	-1	9.56	9.375	0.185

Figure 4.6: Fitted Values and Residuals, Análisis DOE

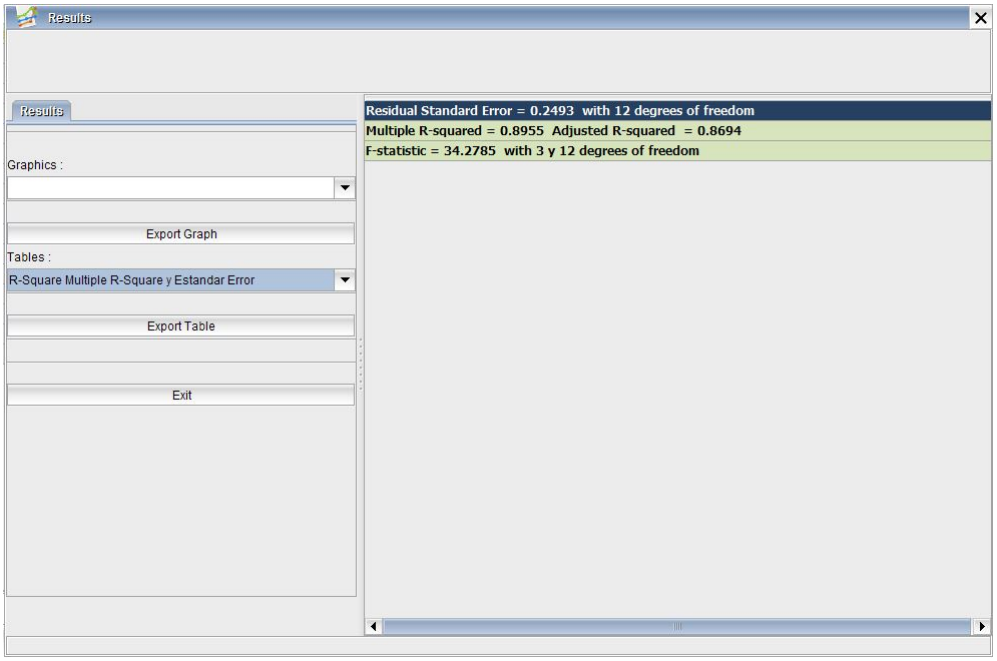
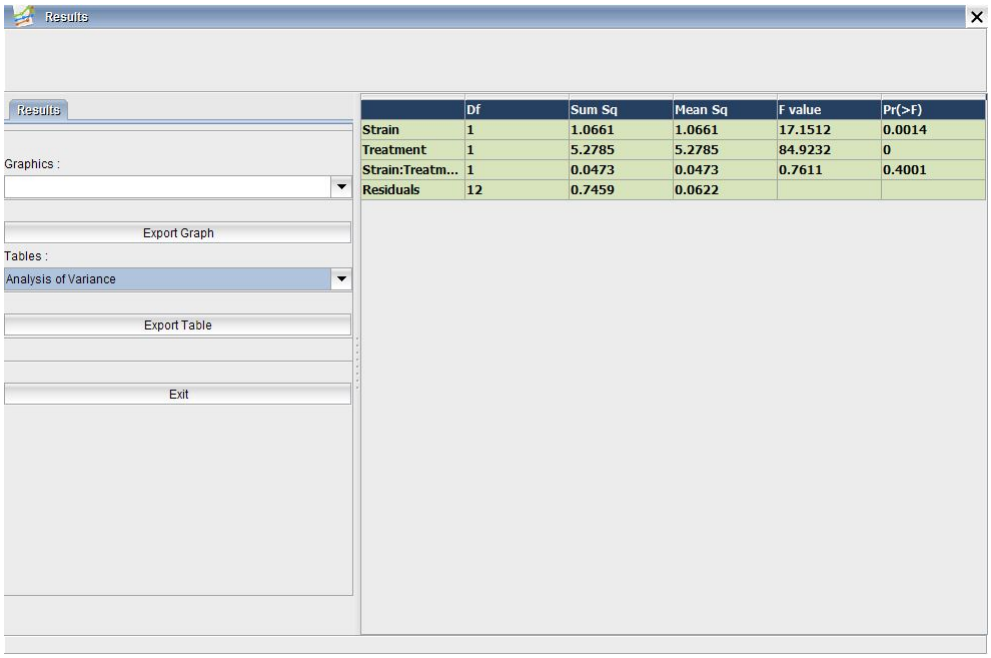

Results
Results
Graphics :
Export Graph
Tables :
R-Square Multiple R-Square y Estandar Error
Export Table
Exit
Residual Standard Error = 0.2493 with 12 degrees of freedom
Multiple R-squared = 0.8955 Adjusted R-squared = 0.8694
F-statistic = 34.2785 with 3 y 12 degrees of freedom

Figure 4.7: R-Squared for model (4.1), DOE Analysis



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Strain	1	1.0661	1.0661	17.1512	0.0014
Treatment	1	5.2785	5.2785	84.9232	0
Strain:Treatm...	1	0.0473	0.0473	0.7611	0.4001
Residuals	12	0.7459	0.0622		

Figure 4.8: ANOVA, DOE Analysis

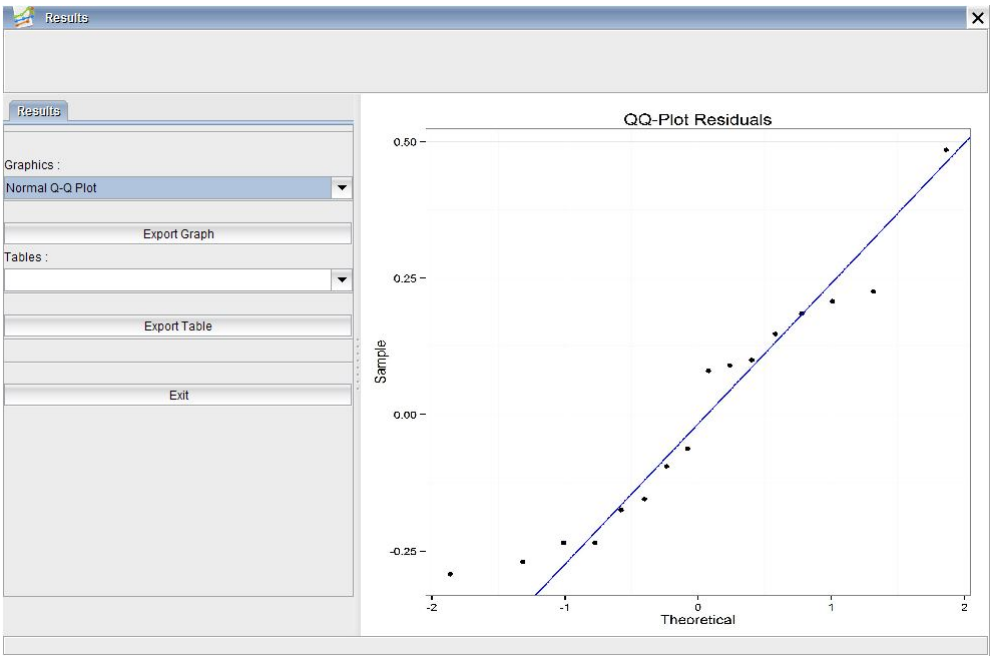


Figure 4.9: Q-Q Plot for Residuals, DOE Analysis

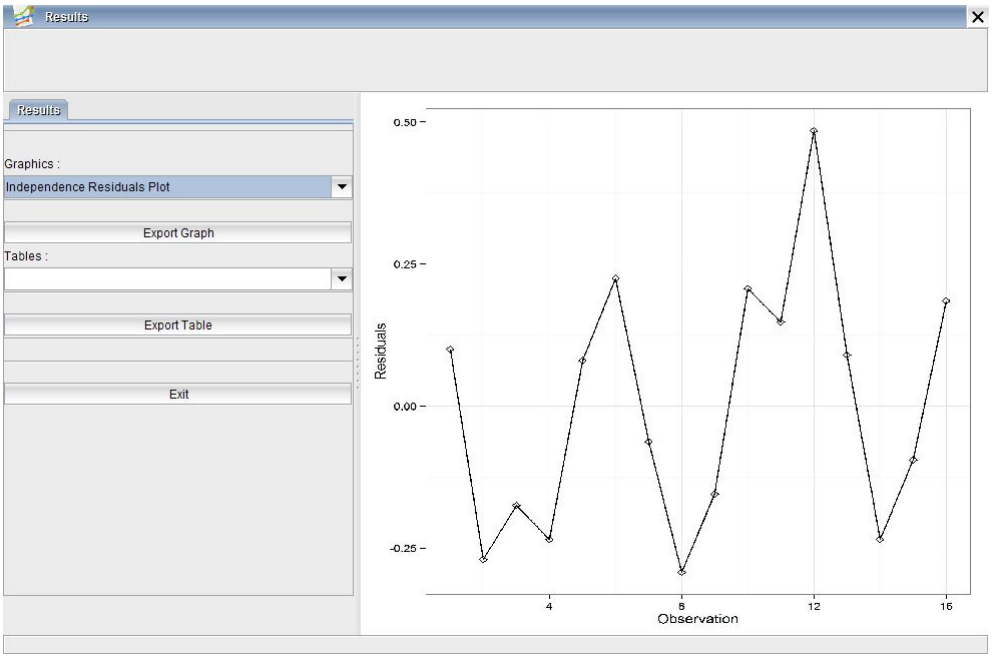


Figure 4.10: Residuals Independence Graph, DOE Analysis

amplitude of the points for each adjusted "level" (value) is the same. If the points on the graph reflect a form of cone, this indicates that the variance of our residuals is not constant. For our data, there is only one point that does not fit in our proposed from. If we take this value as an "outlier" we agree with iii). Figure 4.11.

- **Half-Normal Plot**. Is a graph of the absolute value of estimated factor effects against its normal cumulative probabilities. The Figure 4.12 shows the half-Normal plot of the effects of each of the factors in the design. The solid line of the graph always starts at the origin and usually passes trough the 50% quantile data. This plot is easy to interpret, particularly when there are few effects. For our data, the estimated effects for treatment and strain factors are far away from the midline, so we say these effects can not be due to chance (or do not belong to a normal distribution), and we consider them as significant. The more remote effects are from the midline, the more significant will be on the response. Certainly, this is consistent with the ANOVA and hypothesis testing coefficients shown above.
- **Main Effects Graph**. The plot of the main effects depicts the average of all observations in each of the factor levels and connect them with a line. For a two-level factor, the effect of one factor is related to the slope of the line; the more vertical the line is, the more significant the effect will be. In Figure 4.13 we find that the lines of each of the main are more likely to be vertical lines. For **treatment** we can see a line with a very steep slope down, here we can conclude that treatment has a negative effect on the response. That its, a change from the Control group (-1) to Group therapy (1), is negative and significant. For **strain**, the slope of the line is not very pronounced but not close to being an horizontal line, this means that a change from BALB / c (-1) to C57BL (1), produces a negative effect and this effect is significant but not to a great extent. If the lines of the effects were horizontal (or nearly horizontal), then we would say the estimated effect is not significant since the average of the observations in the Lower level (-1) would be almost equal to the average High level (1); i.e. no change from one level to another.

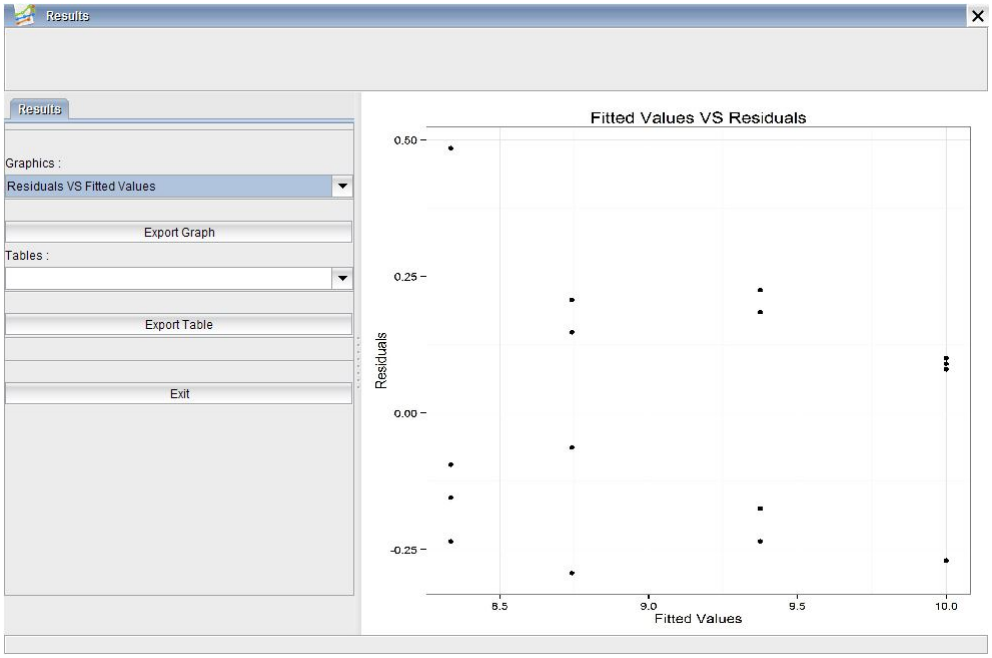


Figure 4.11: Residuals VS Fitted Plot, DOE Analysis

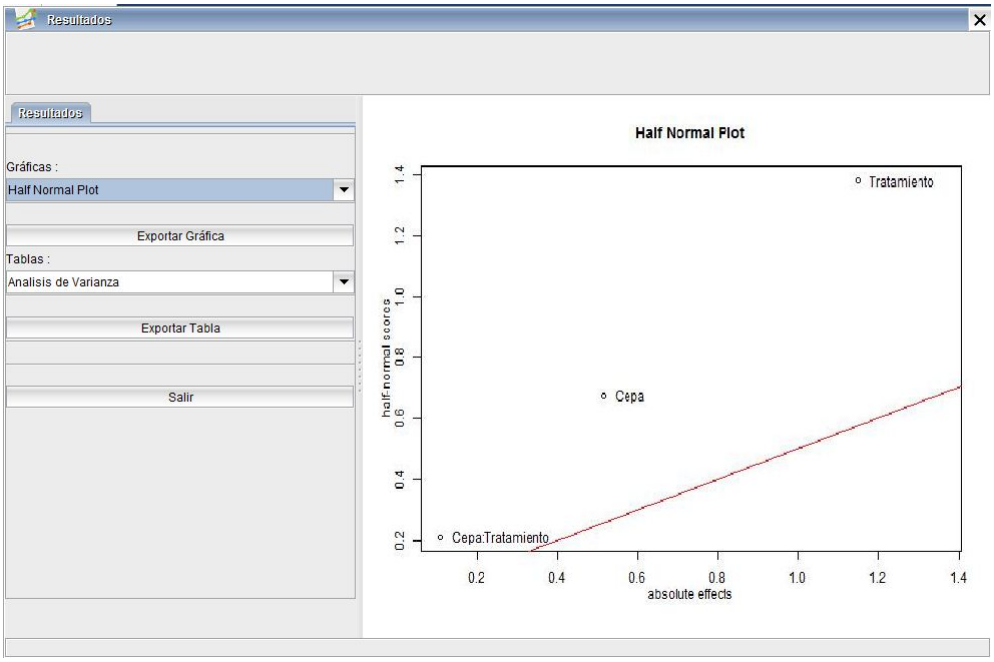


Figure 4.12: Half-Normal Plot, DOE Analysis

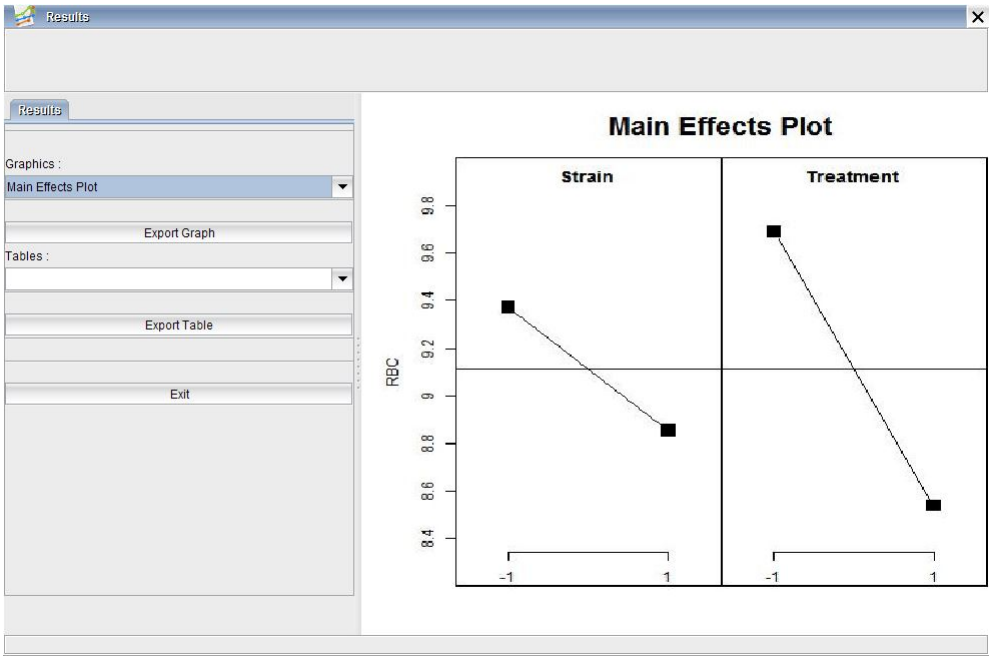


Figure 4.13: Main Effect Plots for RBC experiment, DOE Analysis

- **Interaction Plot**. This graph shows the average of the observations in each of the combinations of two factors A and B [(-, -), (+, -), (-, +), (+, +)], and these averages are joined by a line. For instance, for the lower left graph, the X axis is represented by the **strain** factor and the Y axis is represented by the **Treatment** factor. This chart tells us that if we change the BALB / c strain (-1) to C57BL (1) a decrease occurs in the average of the observations, both for the control group and the treatment group. On the other hand, for the upper right graph, now we have **Treatment** on the X axis and **Strain** on the Y axis. This graph is explained as follows: if we apply the treatment to a rat having the placebo, for any of the two strains, there will be a decrease in average the response. This is an indication that the two factors do not interact. When we say that the graph reflects an interaction between the factors? When changing the factor A, for example, from the level (-1) to level (1), one of the levels in B increases the other decreases the average of the observations.

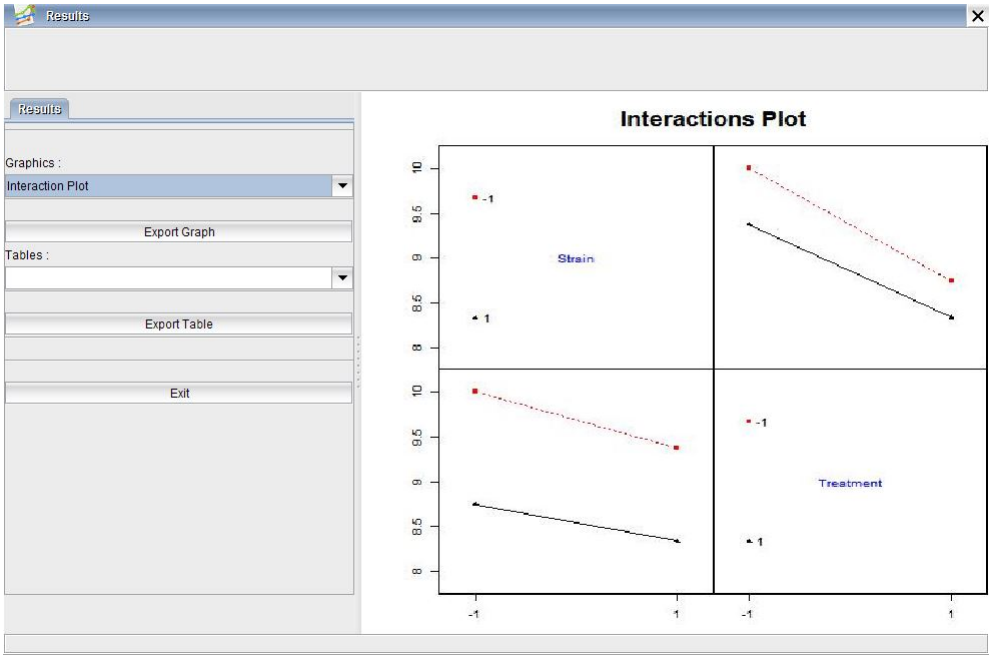


Figure 4.14: Interaction Plot, DOE Analysis

4.2 Longitudinal Analysis

4.2.1 Linear Mixed Model

In longitudinal studies it is well known that temporal observations of a particular subject are correlated. If, for example, we treat 2 patients against flu, the two of them will react in a different way due to their natural differences (genetics, etc). The idea behind the basic analysis of linear mixed models is that regression coefficients of each individual are different. Then, we treat the coefficients as random; i.e. different coefficients for each individual.

The simplest form of a random coefficient model is regarding only the intercept as a random coefficient. This means the first observation of individuals (e.g. before starting the study) are different, but the growth (slope) in the response is the same for everyone. This relationship can be seen in Figure 4.15.

It is also possible that the intercept is the same for all subjects; but growth is different for subjects. This model will include a random slope. This phenomenon is observed in Figure 4.16.

Another case is when both coefficients, the intercept and slope, are random. This means that each individual has his own response "Profile". See Figure 4.17.

Summary of Linear Mixed Effects Model

- The linear mixed effects model can be represented as follows

$$\begin{aligned} y_i &= X_i\beta + Z_i b_i + \varepsilon_i, \quad i = 1,\ldots,M, \\ b_i &\sim N(0,\Sigma), \quad \varepsilon \sim N(0,\sigma^2 I) \end{aligned} \tag{4.2}$$

where β is a p-dimension vector of fixed effects, b_i is a q-dimension vector of random effects, X_i (of size $n_i \times p$) and Z_i (of size $n_i \times q$) regression matrices for fixed and random effects, respectively; and ε_i a n_i-dimension vector of with-in-subjects errors, i.e. the residuals for each observation in time of individual i. An important assumption when

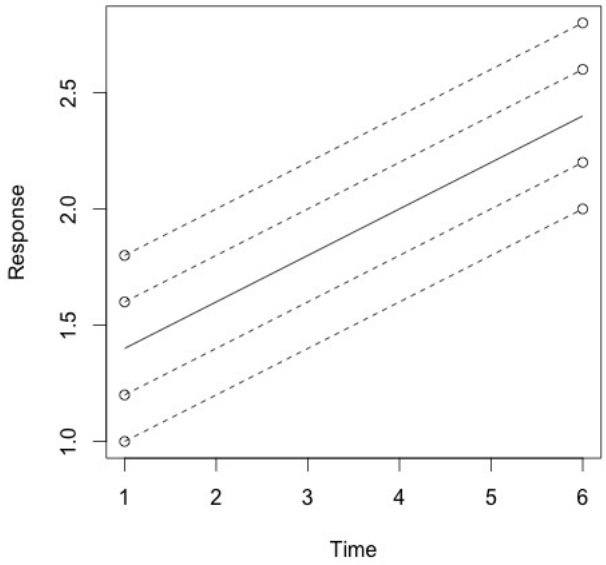


Figure 4.15: Random Intercept Example

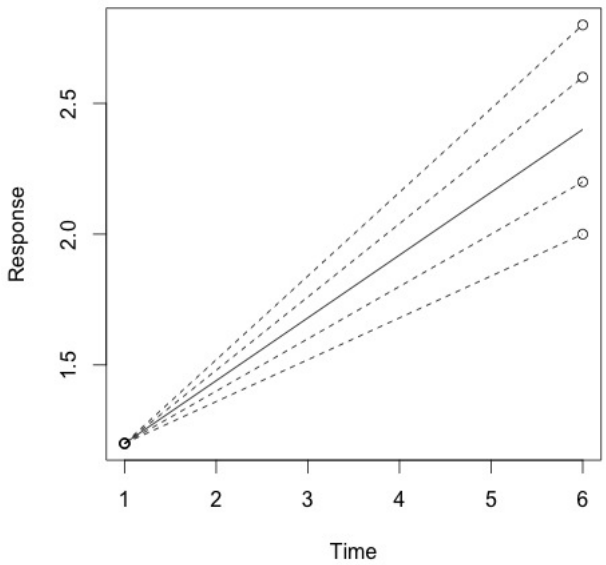


Figure 4.16: Random Slope Example

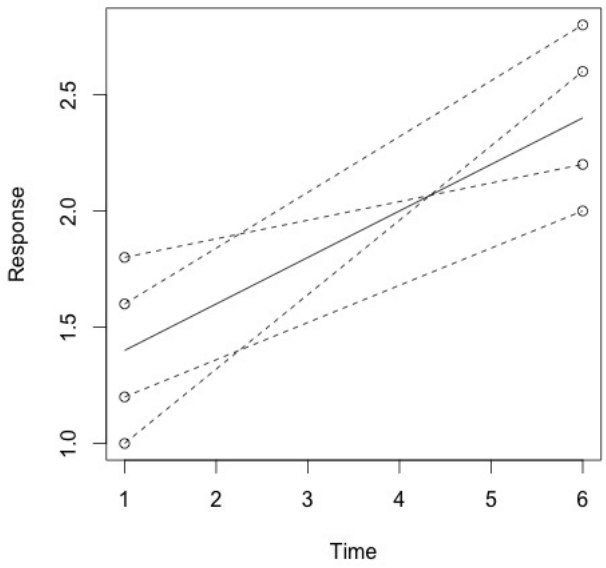


Figure 4.17: Random Intercept and Slope Example

working with linear mixed model is that $Var(\varepsilon_i) = \sigma^2 I$; i.e. the variance for each of the residuals vectors has to be the same. Another assumption of the model is that the the random effects b_i and the residuals ε_i have to be independent. Furthermore, it is also assumed that the random effects follow a multivariate normal with 0 mean and covariance matrix Ψ.

- The maximum likelihood method is used to estimate the parameters β, σ^2 and Σ. Another method of parameter estimation is the **restricted maximum likelihood (REML)** method. This method is the most used, since maximum likelihood tends to inflate the estimates for σ^2 and Σ. We have to keep in mind that the estimation for random coefficients is the covariance matrix Σ; because we are making the assumption that the random effects are Normal distributed. Therefore, in order to give puntual estimations for the random coefficients of each individual we can sample from a Normal distribution with 0 mean and covariance matrix Σ. This is known as **best linear unbiased predictions** for random coefficients.
- In order to test the hypotheses about the significance of the fixed effects we use a conditional t-test. This test is conditioned by the estimates of the random effects, using REML method. There is another test for comparing models based on the likelihoods of the two proposed models. This test basically compares the information explained by two models and decide whether or not our proposed model is better than the other. This test helps us to determine the significance of potential random effects in our data, and is based on a χ^2 distribution. Finally, this type of test can be used using two different criteria: *Akaike Information Criterion* (AIC) and *Bayesian Information Criterion* (BIC). (See A.5)
- The linear mixed effects model (4.2.1) assumes that with-in-subjects errors ε_i are independent and follow a $N(0, \sigma^2 I)$. There is also an extension to linear mixed effects model in which the assumptions are relaxed, and this allows us to model non-constant variance (**heteroskedasticity**) and different correlation structures for errors (see A.3 y A.4 for a

catalog of variances and correlations structures, respectively). This extended model is expressed as follows

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i, \quad i = 1,\ldots,M, \qquad (4.3)$$
$$b_i \sim N(0,\Psi), \quad \varepsilon \sim N(0,\sigma^2\Lambda_i)$$

in the same manner as the basic mixed effect model, we assume that the errors ε_i and the random effects b_i are independent. The estimation methods and hypothesis testing for the parameters of this model are similar to the basic model. The matrix Λ_i is transformed so that we can re-formulate a basic model and apply the known theory.

Linear Mixed Effects Model Analysis

As an example, the data file **weights.csv** contain the weights of 16 mice over time. Eight mice were assigned to a control diet (Diet 1), four to an experimental diet (Diet 2), and the remaining four to other experimental diet (Diet 3). Measurements were made for 64 days after application of the treatment. Measurements were performed every seven days plus one on day 44. We want to analyze the performance of the three diets over time on the weights of the mice. Data loaded in the program is seen in Figure 4.18.

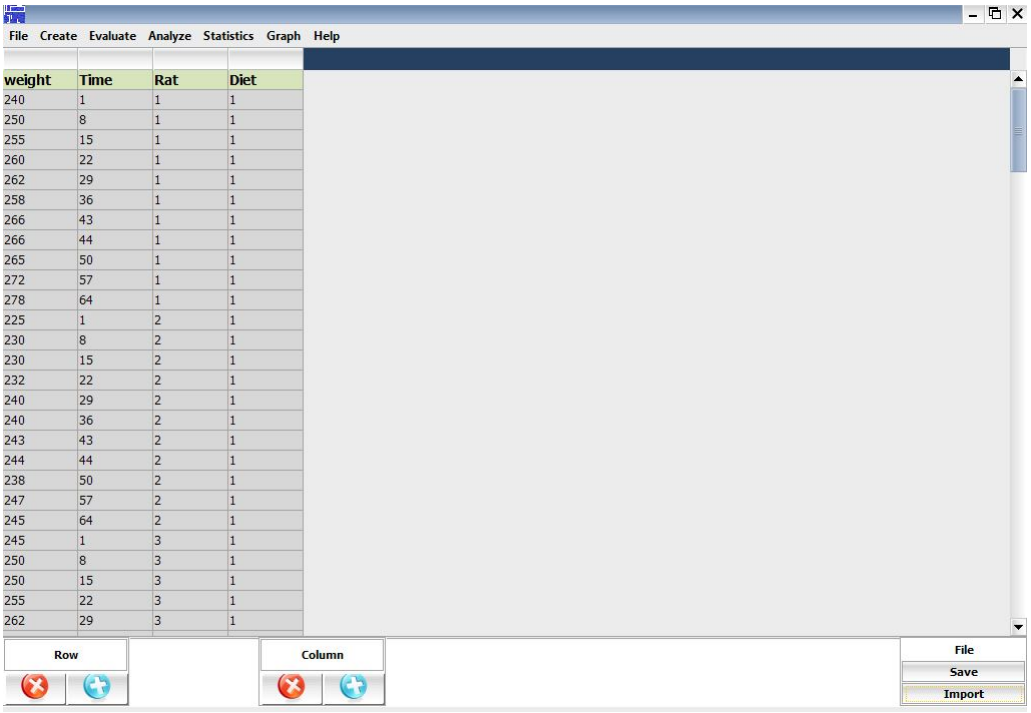


Figure 4.18: Weight Experiment Dataset. File **weights.csv**

To access the functionality of Linear Mixed Models we follow **Analyze** $->$ **Longitudinal Analysis** $->$ **Linear Mixed Models**. In Figure 4.19 we can see differences between individuals belonging to each of the three diets. We can also see an unusual initial weight of the mouse 2 in Diet 2 group. The weights seem to grow linearly in time, but with different initial weights and different slopes for each diet. Therefore, it was decided to use a linear mixed model with random effects in the intercept and time in order to model the mouse-to-mouse variation. This plot as created using the **Longitudinal Graph** function (Section 6.1).

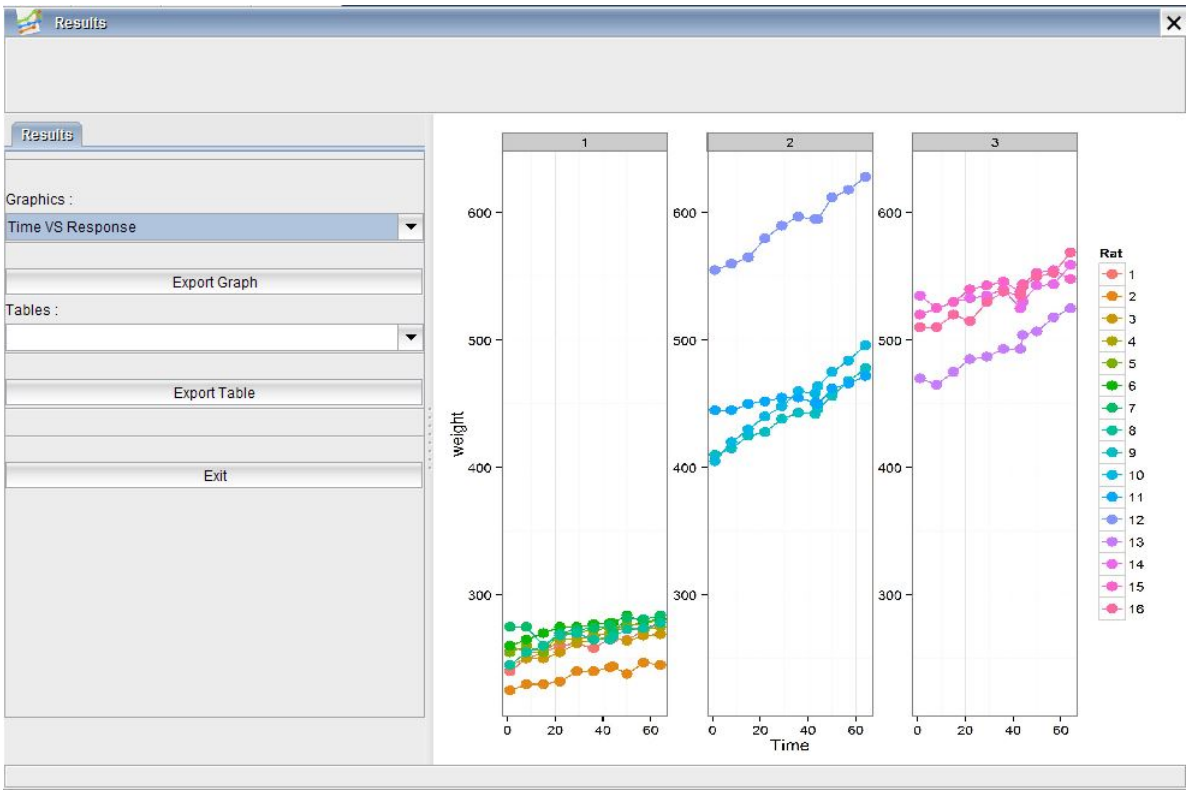


Figure 4.19: Longitudinal (trellis) plot Weights Experiment

The mixed linear model for this analysis can be expressed as follows

$$Y_{it} = (\beta_0 + \gamma_{02}D_{2i} + \gamma_{03}D_{3i} + b_{0i}) + (\beta_1 + \gamma_{12}D_{2i} + \gamma_{13}D_{3i} + b_{1i})t_{ij} + \varepsilon_{it} \tag{4.4}$$

where D_{2i} is a binary variable that takes the value of 1 if the ith mouse recieves Diet 2; D_{3i} is a binary indicator for variable Diet 3, i.e. 1 if the mouse received Diet 3; β_0 y β_1 are, respectively, the random intercept and slope under Diet 1. In this case, Diet 1 is called "reference", because our results will be comparison against it. γ_{0k} is the average difference of the intercept under Diet k minus the intercept under Diet 1; γ_{1k} is the average difference of the slope under Diet k (for our example, k = 2, 3) and the slope under Diet 1; b_{0i} is the random intercept, i.e. the particular mouse initial weight; b_{1i} is the random slope, i.e. the particular mouse growth over time. Finally, e_{ij} is the model error. This is an example of a model using **Treatment** contrasts in our model matrix. The model implemented in LADES can be seen in Figure 4.20, where we can see the main screen of the function **Linear Mixed Model**.

After clicking **Adjust**, the function displays different outputs such as ANOVA, parameter estimation etc.; but, for now we will concentrate on the **Adjusted Vs. Residuals** plot. This graph of standardized residuals versus fitted values gives us a clear indication of the heteroskedastisity in the residuals. In Figure, we can see that our residuals indicate heteroskedastisity. Thus, in order to have residuals with constant variance, we have to use a **Power Variance Structure** . This customization in our analysis can be done using the **Customize** button on the screen in Figure 4.20. The customization menu shown in Figure 4.22.

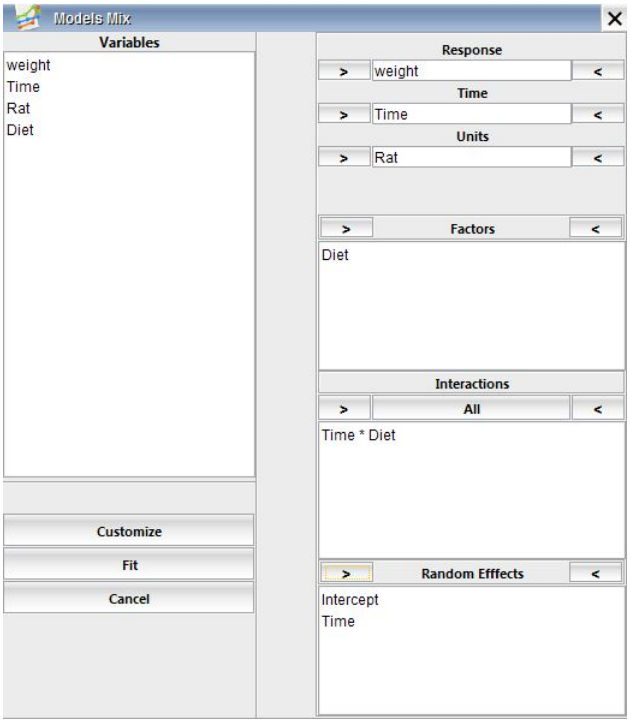


Figure 4.20: Main Screen, Linear Mixed Models

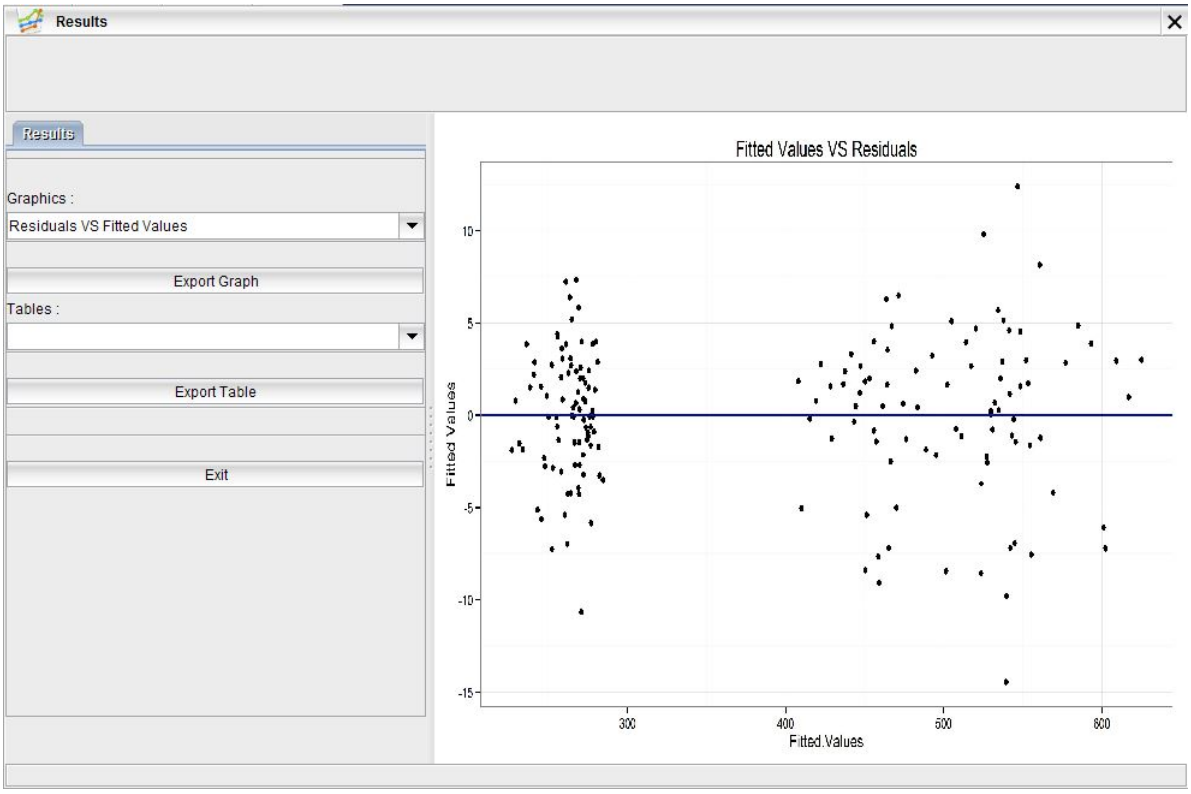


Figure 4.21: Results of longitudinal graph

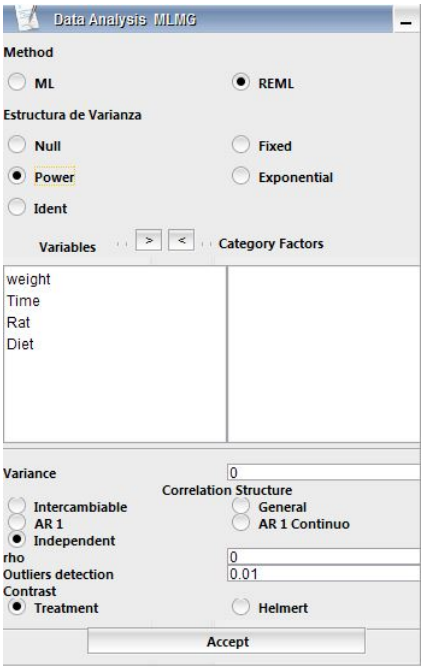


Figure 4.22: Customization Screen, Linear Mixed Models

Function Outputs

- **95 % Confidence Intervals for Fixed Parameter Estimates.** Figure **??** shows 95% confidence intervals and estimates for the parameter in model (4.2.4). In addition, confidence intervals for correlation and variance structure parameters are shown. A first evaluation to determine which factors are significant is that its confidence interval does not include zero .
- **95 % Confidence Intervals for Mixed Effects Parameter Estimates.** Figure 4.24 shows 95% confidence intervals and estimates for each mixed effect parameter included in the model.
- **95 % Confidence Intervals for Variance Error Estimate.** Figure 4.25 shows 95% confidence interval and estimate for the variance of the error.
- **Likelihood Criteria.** In this section, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) (see A.5) are displayed. In addition, the log-likelihood is presented. See Figure 4.26.
- **Residuals Statistics.** In this section, see Figure 4.27, some statistics about the residuals of the model are shown. We note the median is closed to 0, and the quartile 1 is closed to the minimum value, same for Quartile 3 and the maximum. As a result, this gives evidence residuals may follow a symmetric distribution around zero.
- **Fixed Effects Correlations.** In Figure 4.28, we observe the correlation matrix of the fixed effects estimates.
- **Adjusted Values and Residuals.** In Figure 4.29, the adjusted values and residuals of model (4.2.1) are shown. We note that the adjusted values are similar to the values of the original observations. We can say that our model provides a good representation of the data.
- **Analysis of Variance.** This section shows the results of the ANOVA. The sum of squares is shown for each of the dependent variables, as well as the construction of each of the F-statistic. Note all the fixed effects are significant. Therefore, there is a significant change

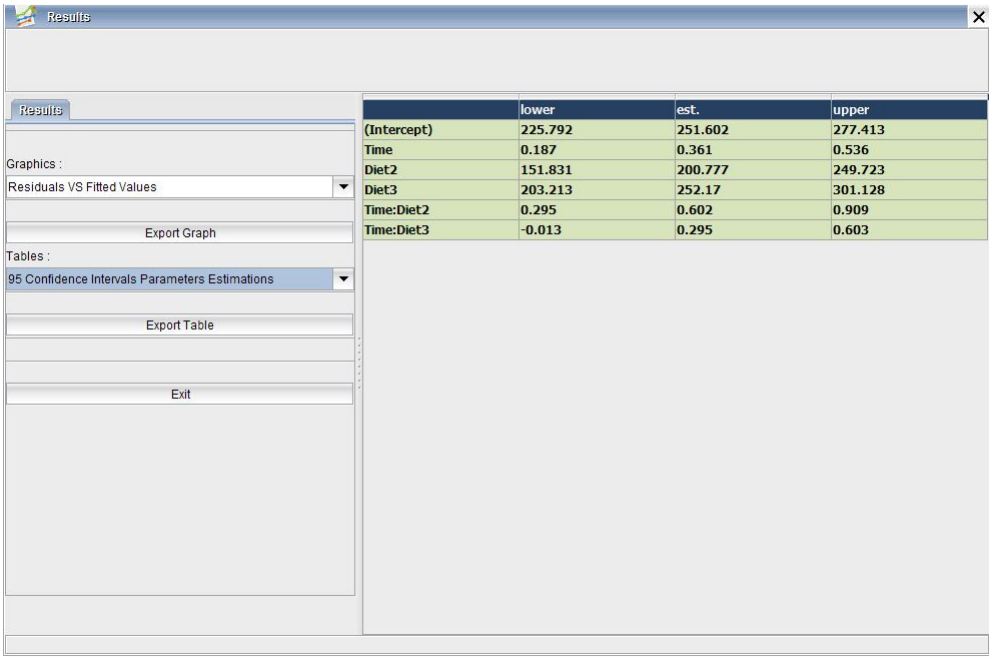


Figure 4.23: Confidence Intervals of Parameter Estimates, Linear Mixed Models

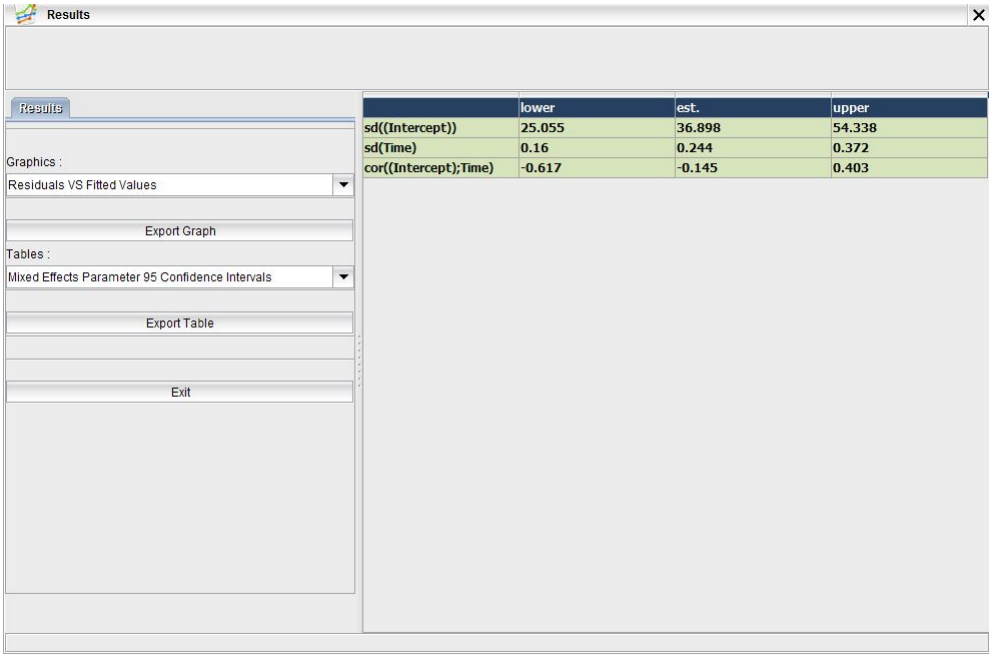


Figure 4.24: Confidence Intervals for Mixed Effects Parameter Estimates, Linear Mixed Model

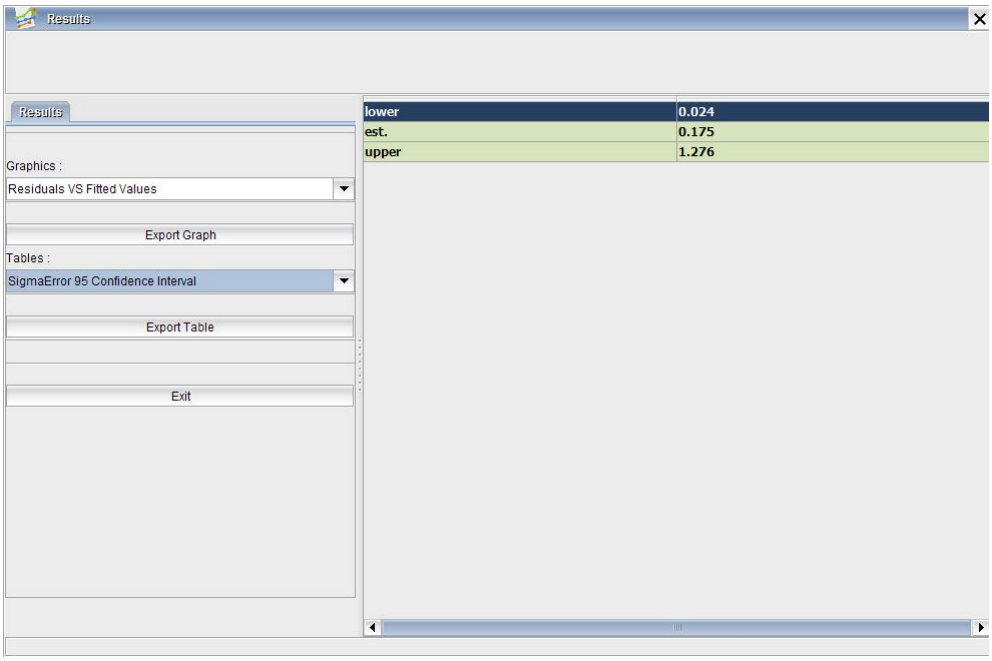


Figure 4.25: Confidence Intervals for Variance Error Estimate, Linear Mixed Model

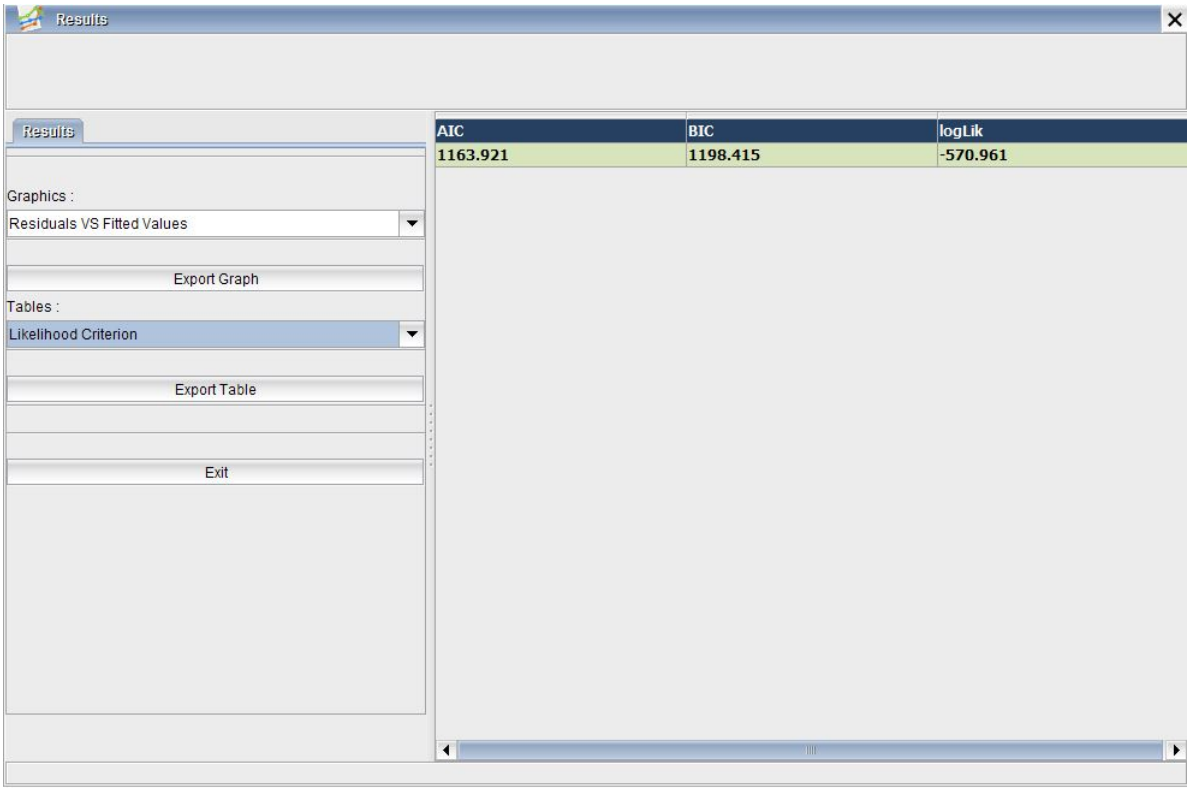


Figure 4.26: Likelihood Criteria, Linear Mixed Models

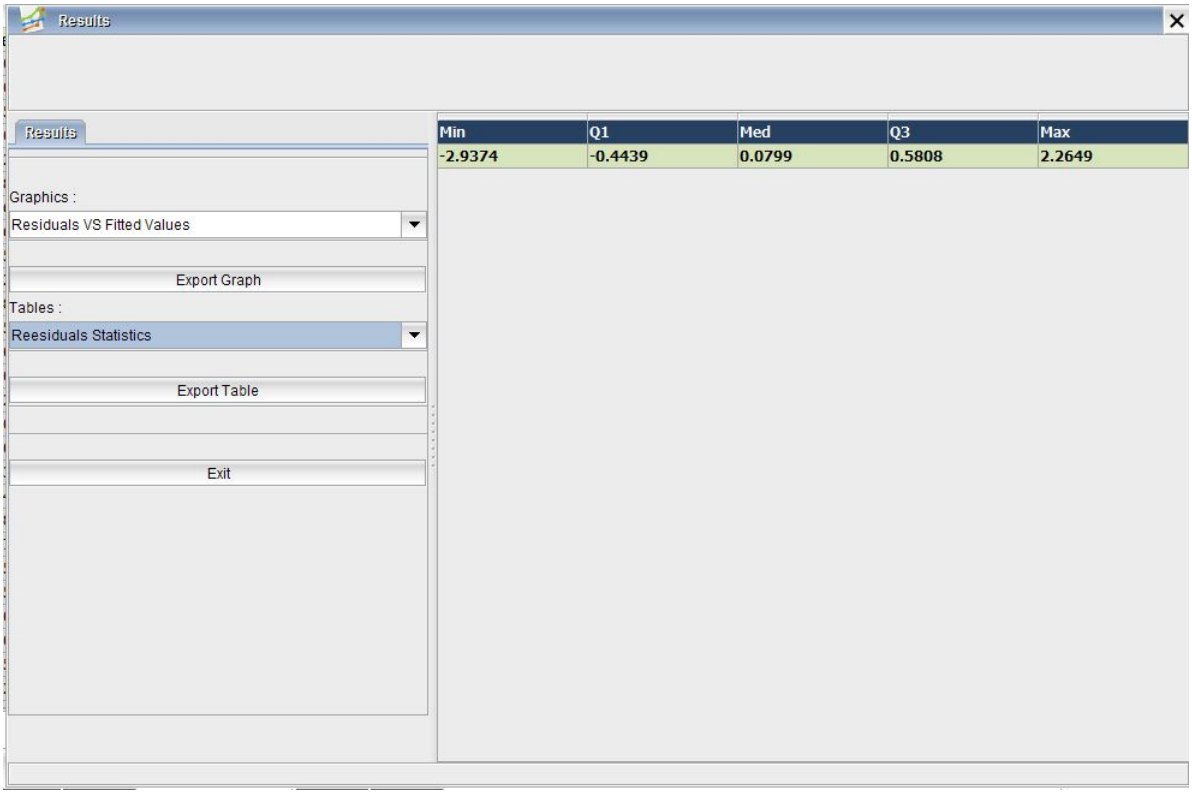


Figure 4.27: Residuals Statistics, Linear Mixed Models

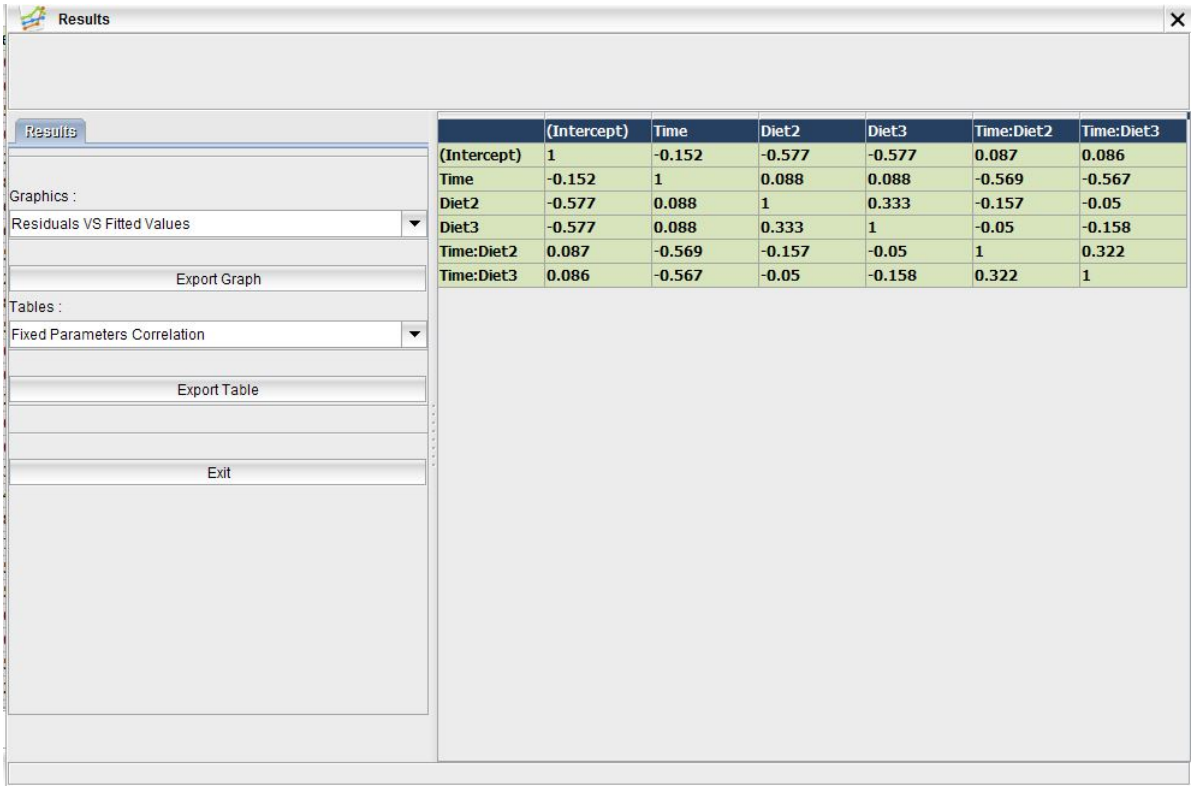
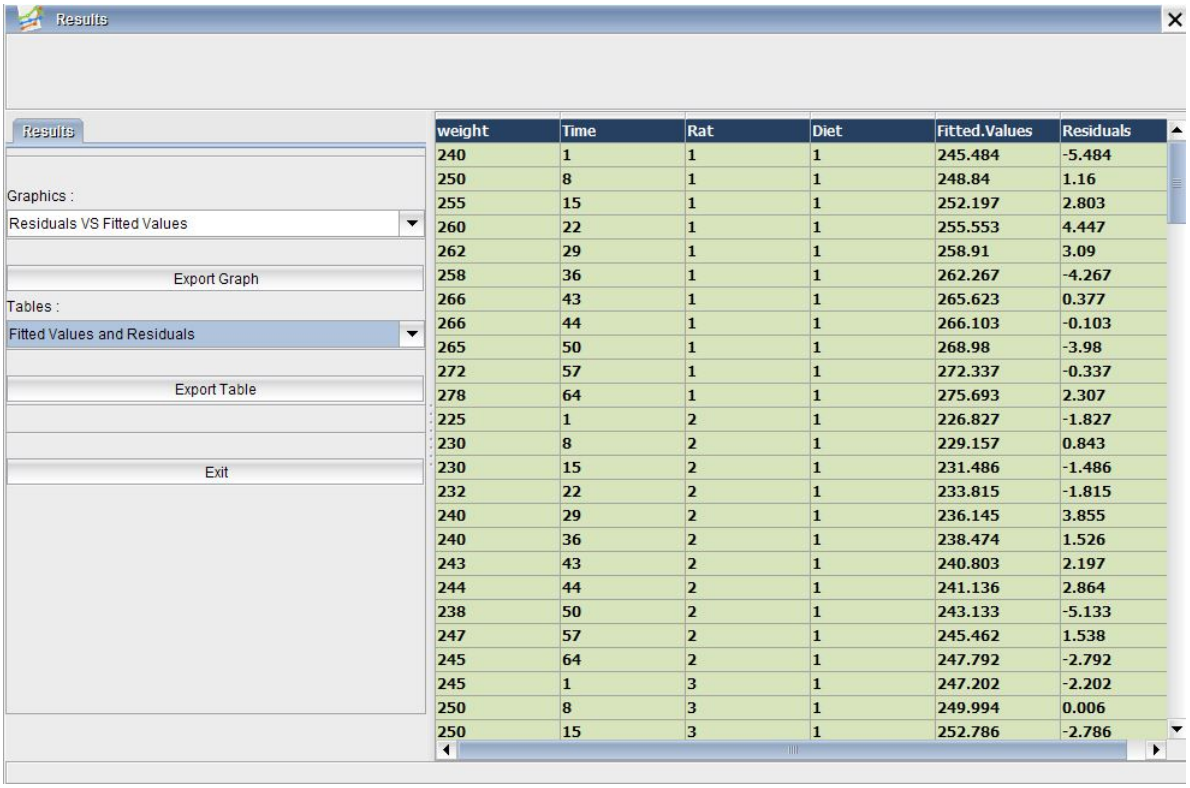

Results
Results
Graphics :
Residuals VS Fitted Values
Export Graph
Tables :
Fixed Parameters Correlation
Export Table
Exit
(Intercept) Time Diet2 Diet3 Time:Diet2 Time:Diet3
(Intercept) 1 -0.152 -0.577 -0.577 0.087 0.086
Time -0.152 1 0.088 0.088 -0.569 -0.567
Diet2 -0.577 0.088 1 0.333 -0.157 -0.05
Diet3 -0.577 0.088 0.333 1 -0.05 -0.158
Time:Diet2 0.087 -0.569 -0.157 -0.05 1 0.322
Time:Diet3 0.086 -0.567 -0.05 -0.158 0.322 1

Figure 4.28: Fixed Effects Correlations, Linear Mixed Models

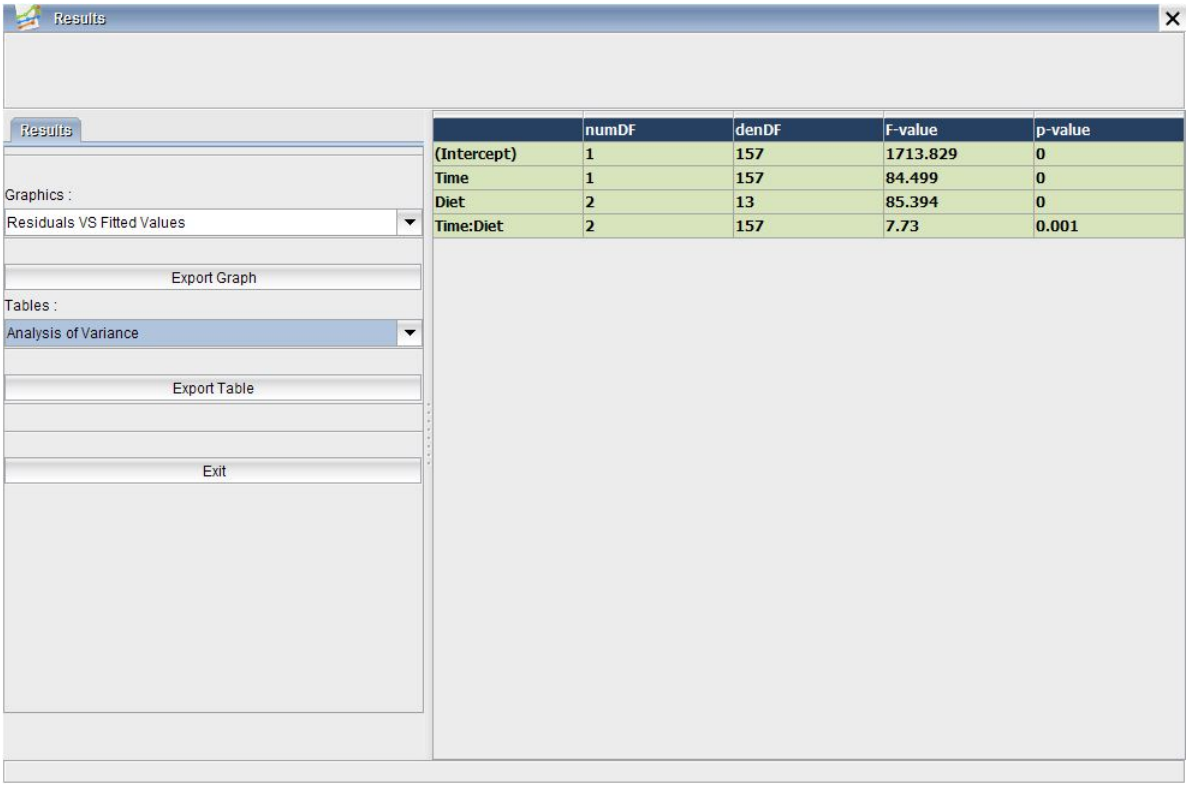


weight	Time	Rat	Diet	Fitted.Values	Residuals
240	1	1	1	245.484	-5.484
250	8	1	1	248.84	1.16
255	15	1	1	252.197	2.803
260	22	1	1	255.553	4.447
262	29	1	1	258.91	3.09
258	36	1	1	262.267	-4.267
266	43	1	1	265.623	0.377
266	44	1	1	266.103	-0.103
265	50	1	1	268.98	-3.98
272	57	1	1	272.337	-0.337
278	64	1	1	275.693	2.307
225	1	2	1	226.827	-1.827
230	8	2	1	229.157	0.843
230	15	2	1	231.486	-1.486
232	22	2	1	233.815	-1.815
240	29	2	1	236.145	3.855
240	36	2	1	238.474	1.526
243	43	2	1	240.803	2.197
244	44	2	1	241.136	2.864
238	50	2	1	243.133	-5.133
247	57	2	1	245.462	1.538
245	64	2	1	247.792	-2.792
245	1	3	1	247.202	-2.202
250	8	3	1	249.994	0.006
250	15	3	1	252.786	-2.786

Figure 4.29: Adjusted Values and Residuals, Linear Mixed Models

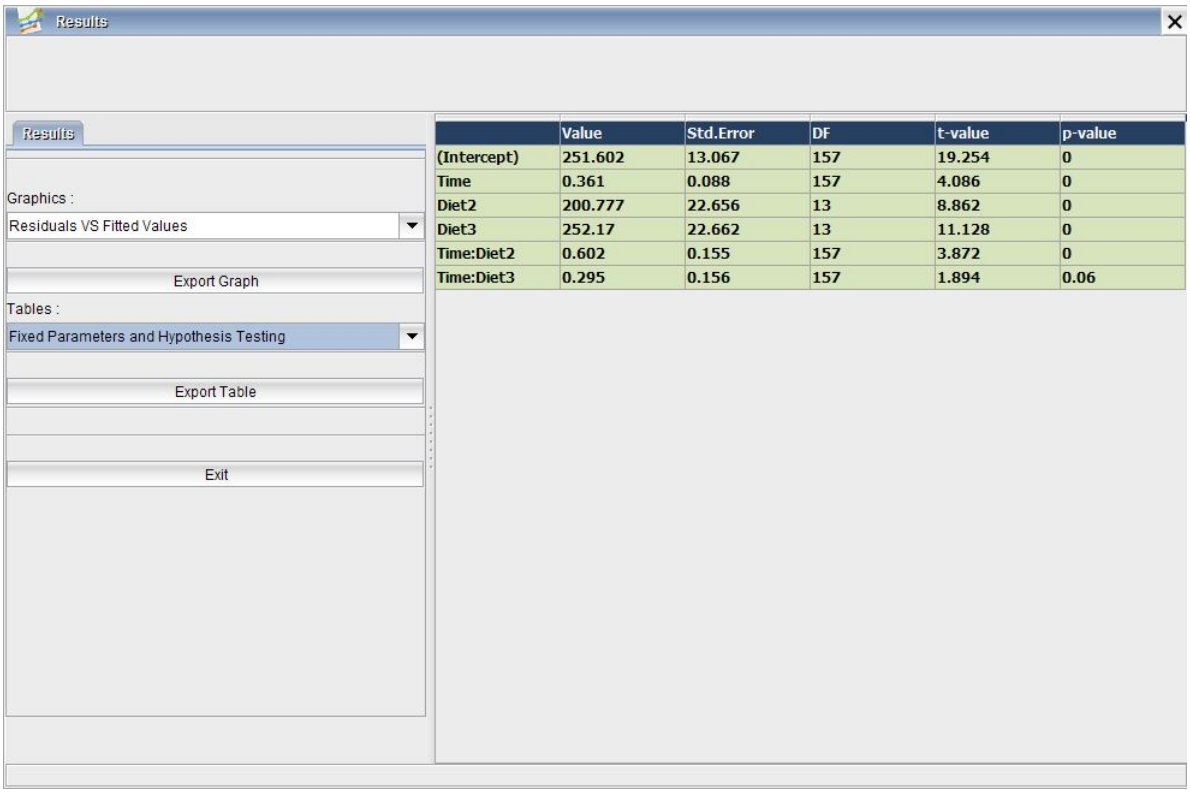
in weight during time, and that change depends on the type of Diet (see Figure 4.30).

- **Fixed Effects and Hypothesis Testing.** For each of the fixed factors included in the model, the regression coefficient, the standard error coefficient and the corresponding p-value are given. The p-value is based on the Wald statistic, which is defined as the ratio between the squared regression coefficient and its standard error. This statistic follows a χ^2 distribution with 1 degree of freedom. We can see that **Time** has an impact of 2.196 gr of weight per unit time. And this change is statistically significant (at $\alpha = 0.05$). Since the contrasts were defined as *Treatment* the Diet factor was splitted into two effects: **Diet 2** and **Diet 3**. Thus, the reference now is Diet 1. Then, the estimated coefficient for Diet 2 refers to the added value to apply Diet 2 in an individual taking Diet 1. Such estimate is significant and this added value is 198,584 grams. Furthermore, for Diet 3 the estimated coefficient is the contribution of the Diet 3 to Diet 1, which produces an increase of 250.809 gr. This estimation is also statistically significant. Regarding the interaction **Time:Dieta2**, it refers to the change in the slope produced by applying Diet 2 to an individual taking Diet 1. This interaction is significant and causes a change of 3.73 grams per unit time. Interaction **Tiempo:Dieta3** is explained likewise; but, this interaction is not significant.
- **Random Effects Estimates by Subject.** In Figure 4.32, the best linear unbiased predictors for random effects are shown each of the rats. These predictions were used to calculate the residuals and fitted values of the model. Basically, the best linear unbiased predictions denote the deviations of each individuals from the **Intercept** (average) and the estimated **Time** effect (average change rate).
- **Estimated Parameters by Subject.** Figure 4.33 shows all the estimated coefficient and best linear unbiased predictions for the fixed and random effects, respectively.



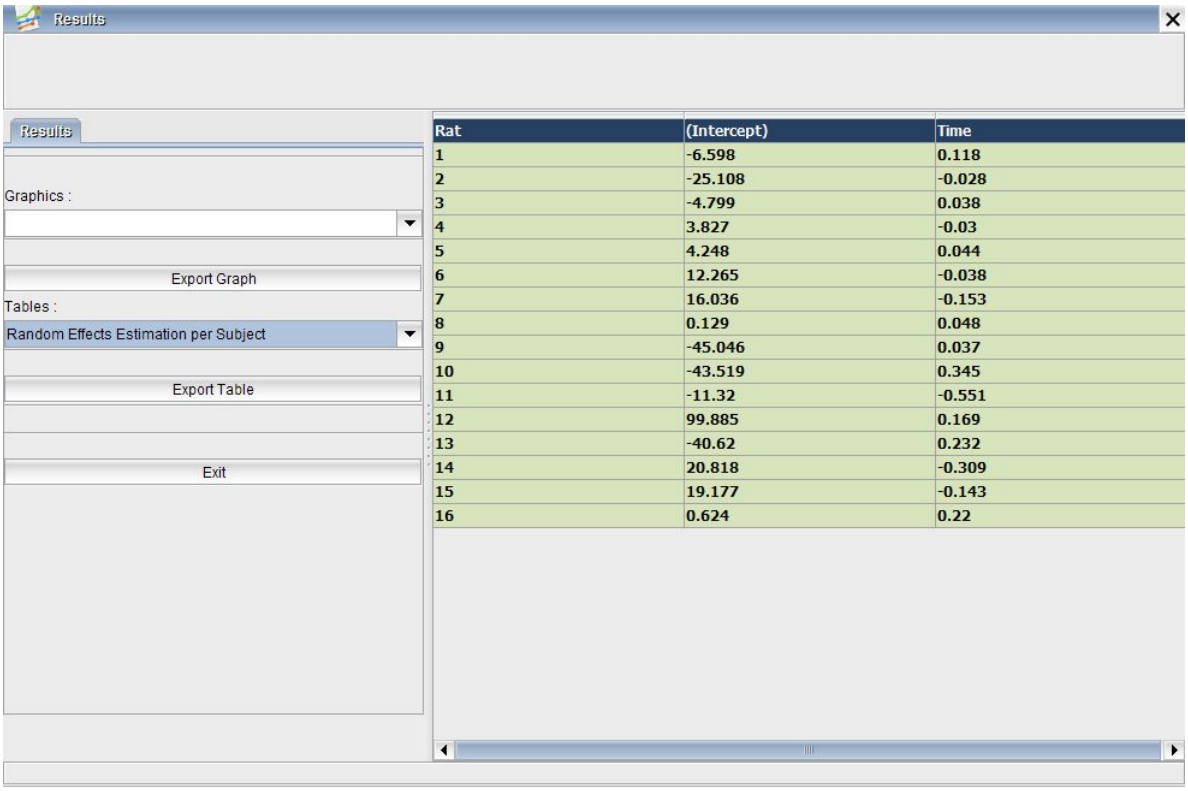
	numDF	denDF	F-value	p-value
(Intercept)	1	157	1713.829	0
Time	1	157	84.499	0
Diet	2	13	85.394	0
Time:Diet	2	157	7.73	0.001

Figure 4.30: Analysis of Variance, Linear Mixed Models



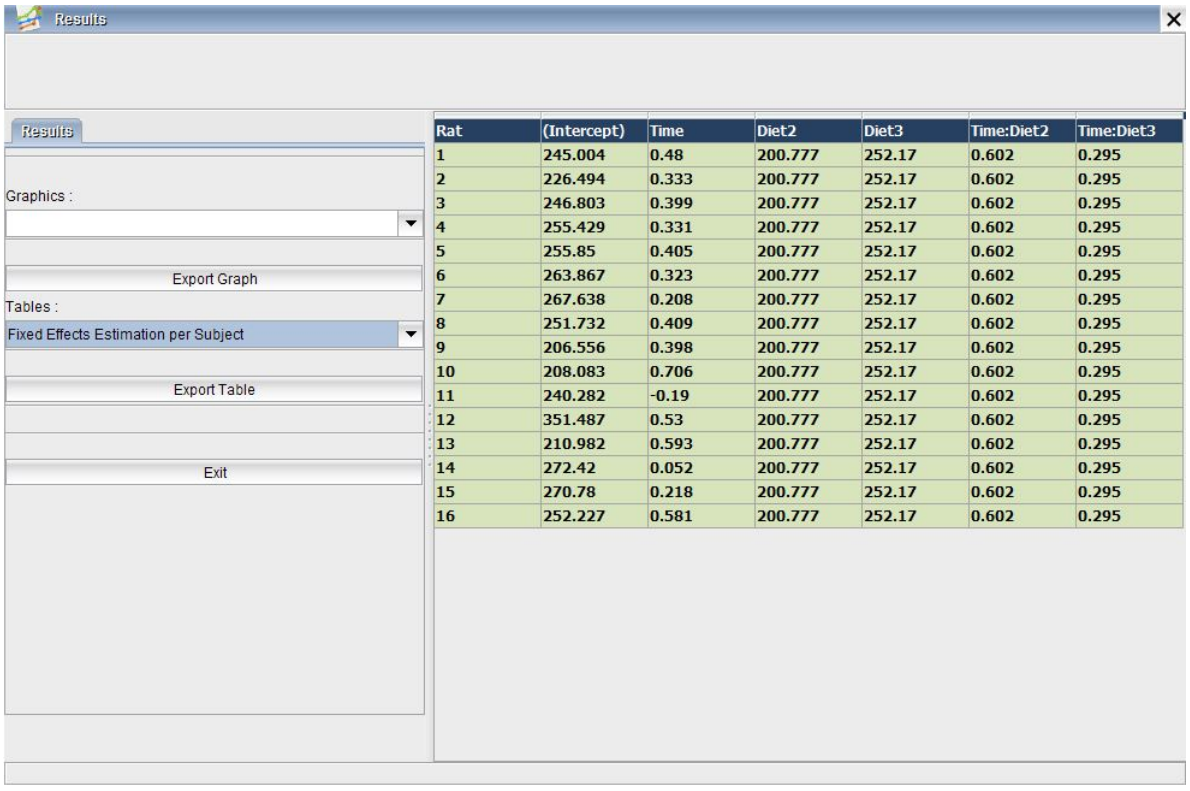
	Value	Std.Error	DF	t-value	p-value
(Intercept)	251.602	13.067	157	19.254	0
Time	0.361	0.088	157	4.086	0
Diet2	200.777	22.656	13	8.862	0
Diet3	252.17	22.662	13	11.128	0
Time:Diet2	0.602	0.155	157	3.872	0
Time:Diet3	0.295	0.156	157	1.894	0.06

Figure 4.31: Fixed Effects and Hypothesis Testing, Linear Mixed Models



Rat	(Intercept)	Time
1	-6.598	0.118
2	-25.108	-0.028
3	-4.799	0.038
4	3.827	-0.03
5	4.248	0.044
6	12.265	-0.038
7	16.036	-0.153
8	0.129	0.048
9	-45.046	0.037
10	-43.519	0.345
11	-11.32	-0.551
12	99.885	0.169
13	-40.62	0.232
14	20.818	-0.309
15	19.177	-0.143
16	0.624	0.22

Figure 4.32: Random Effects Estimates by Subject, Lineal Mixed Effects



Rat	(Intercept)	Time	Diet2	Diet3	Time:Diet2	Time:Diet3
1	245.004	0.48	200.777	252.17	0.602	0.295
2	226.494	0.333	200.777	252.17	0.602	0.295
3	246.803	0.399	200.777	252.17	0.602	0.295
4	255.429	0.331	200.777	252.17	0.602	0.295
5	255.85	0.405	200.777	252.17	0.602	0.295
6	263.867	0.323	200.777	252.17	0.602	0.295
7	267.638	0.208	200.777	252.17	0.602	0.295
8	251.732	0.409	200.777	252.17	0.602	0.295
9	206.556	0.398	200.777	252.17	0.602	0.295
10	208.083	0.706	200.777	252.17	0.602	0.295
11	240.282	-0.19	200.777	252.17	0.602	0.295
12	351.487	0.53	200.777	252.17	0.602	0.295
13	210.982	0.593	200.777	252.17	0.602	0.295
14	272.42	0.052	200.777	252.17	0.602	0.295
15	270.78	0.218	200.777	252.17	0.602	0.295
16	252.227	0.581	200.777	252.17	0.602	0.295

Figure 4.33: Estimated Parameters by Subject, Linear Mixed Models

- **Residuals VS Fitted Values Plot** In this graph (Figure 4.34) we can validate the assumption of constant variance for the residuals. The adjusted values are on the X-axis and the standardized residuals on the Y axis. The graph should show a pattern of an "horizontal band". If the graph is cone shaped, then we say that the variance of the residuals is not constant. In our case, the graph if represents the pattern of constant variance in the residuals. This is the result of including a Power variance structure mentioned at the beginning of the analysis.

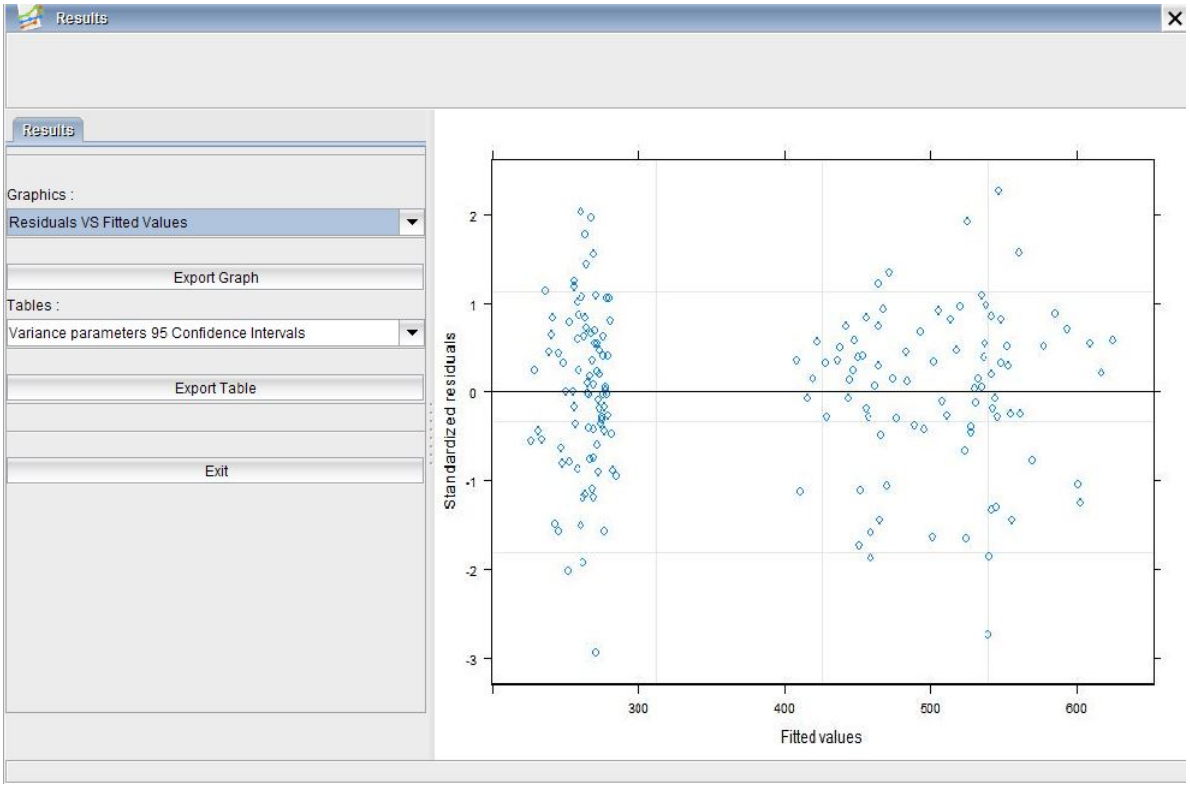


Figure 4.34: Residuals VS Fitted Values Plot, Linear Mixed Models

- **Residuals VS Fitted Values Plot by Factor.** This graph shows the previous graph but split it by Diet. We note that the dispersion for Diet 2 points is smaller than for Diets 1 and 3 (see Figure 4.35). This makes us doubt of the validity of the assumption of constant variance in the residuals, and makes us think that the problem is in Diet 2 observations.
- **Normal Random Effects Plot.** This graph (Figure 4.36) helps us to validate the assumption that the random effects follow a normal distribution. Quantiles for the random effects (X-axis) are shown, and the Normal Quantiles (Y axis). The points for each of the random effects should form a diagonal. For our data, this is true, but we can detect some outliers (rat number 12).
- **Residuals by Subject Plot.** This graph shows the residuals for each one of the 16 subjects. Another assumption of the model is that the variance is the same for all subjects. In Figure 4.37, we note that the dispersion of points for mouse 14 is wider than the others, so that this assumptions could not be fulfilled.
- **Real VS Fitted Values.** In this plot the real versus the fitted values are depicted, see Figure 4.38. This graph help us with detecting discrepancies of the model predictions.
- **Random Effects by Factor Plot** In this section of results we can observe the random Intercept effect on X axis and Time on Y axis for each diet (as this is our factor interest).

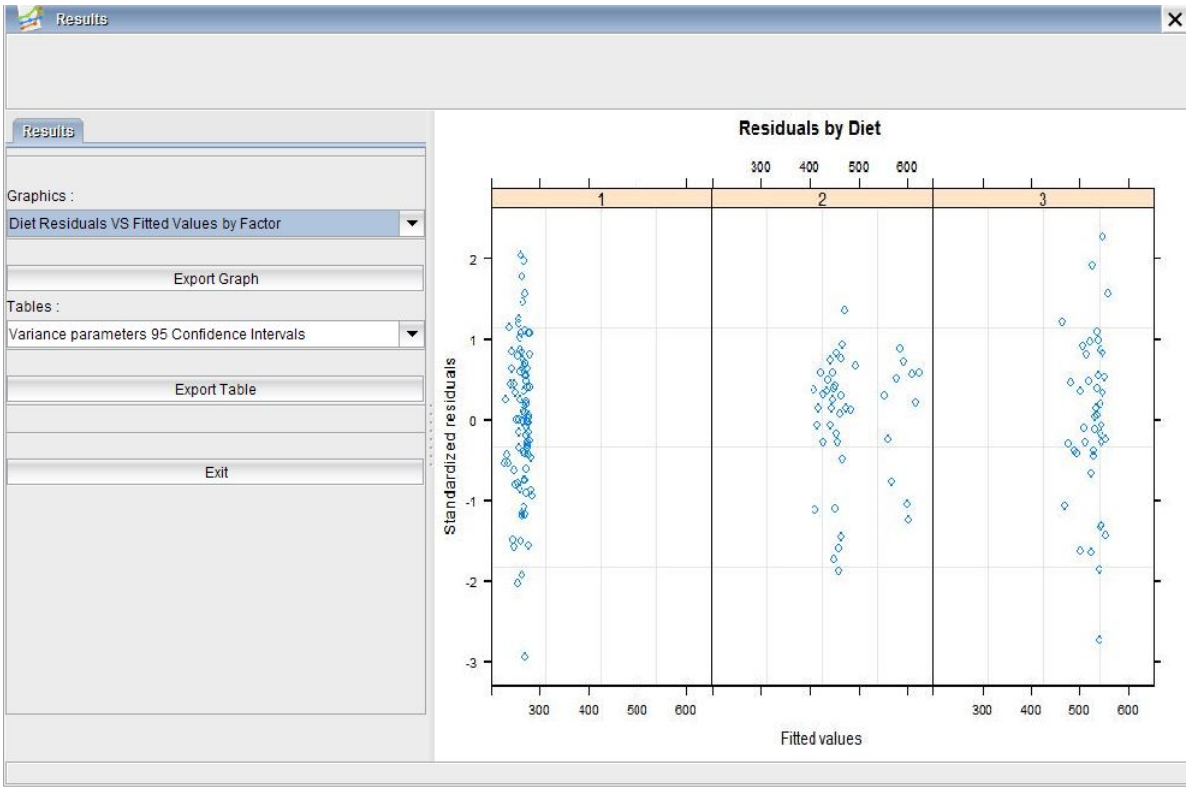


Figure 4.35: Residuals VS Fitted Values Plot by Factor, Linear Mixed Models

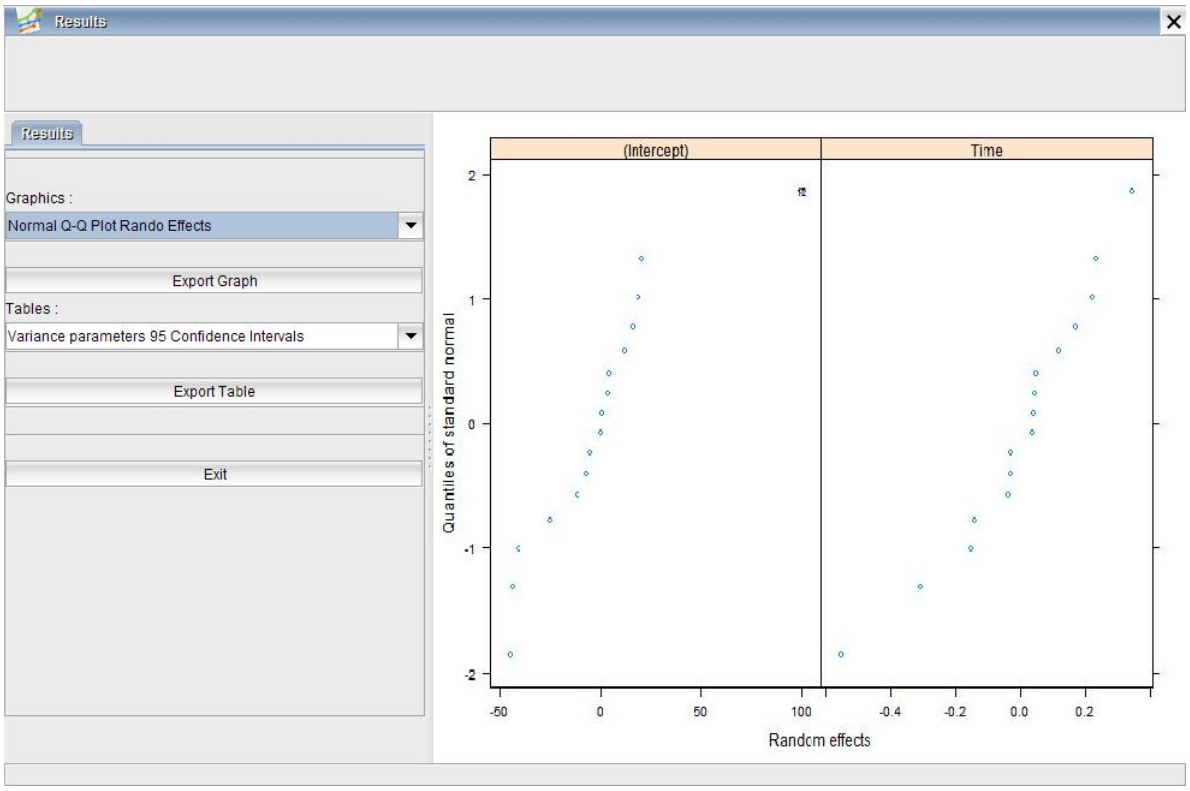


Figure 4.36: Gráfica de normalidad de efectos aleatorios, Modelos Mixtos

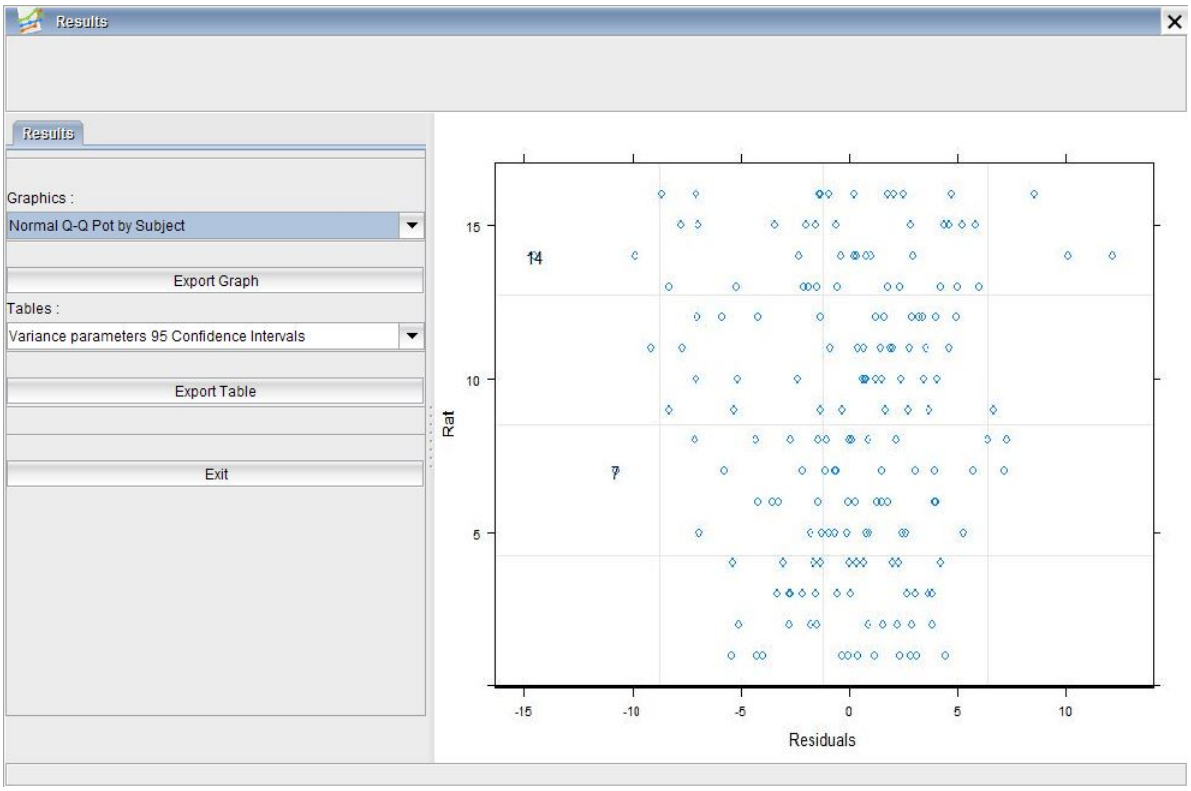


Figure 4.37: Residuals by Subject Plot, Linear Mixed Models

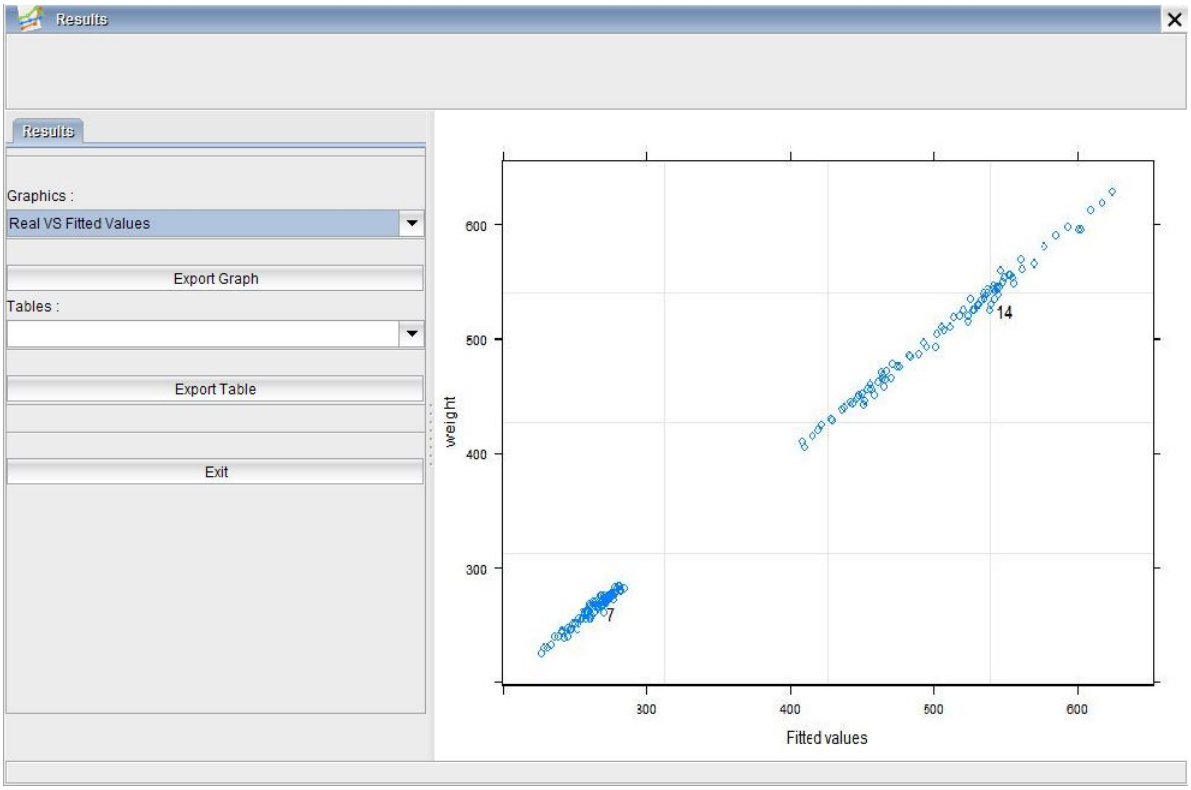


Figure 4.38: Real VS Fitted Values Plot, Linear Mixed Models

These graphs should not present trends as this would let us think about a correlation between in random effects. As mentioned before, other assumptions of the model is that the random effects are uncorrelated. For our case, there were no trends, see Figure 4.39.

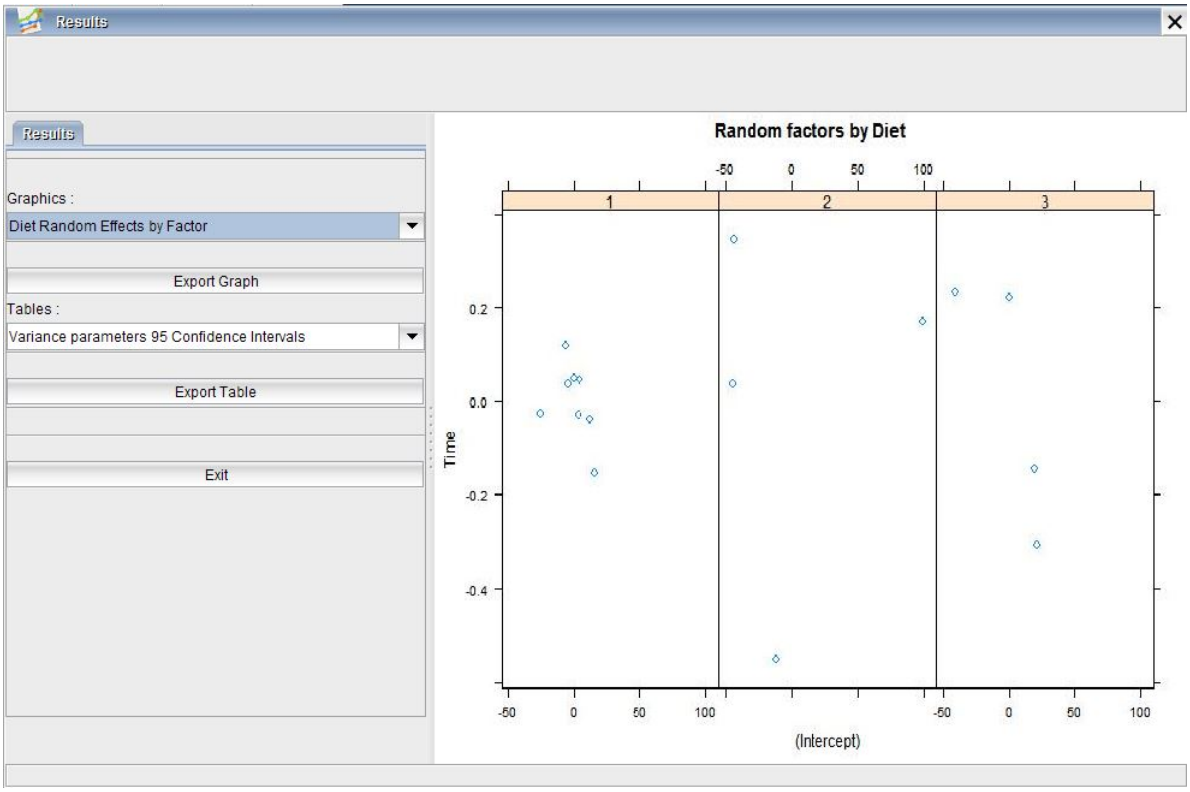


Figure 4.39: Gráfica de efectos aleatorios por factor, Modelos Mixtos

4.2.2 Generalized Linear Mixed Models

When the linear mixed model (LMM) is used, it is assumed that the response of interest is continuous. Sometimes when the variable of interest is ordinal, LMM could be used as long as the response has a large number levels. But LMM does not work to model binary responses or ordinal responses with few levels; or when the response represents counts. For these cases we used generalized linear mixed models (GLMM). This is simply an extension of linear mixed models, that works for more general cases.

Summary of Generalized Linear Mixed Models

- Linear mixed model is used to model the average of observations since it works with continuous responses. This average is said to be a conditional on the random effects, i.e. $y|u$, where u are the random effects. This random variable follows a distribution $N(\mu_{y|u}, \sigma^2 I)$, when y is a continuous variable. Then, the model can be written as follows: $\mu_{y|u}(u) = X\beta + Zb$. Where u is the random effect coefficients; β is the fixed effects vector; X is the model matrix for fixed effects; and Z is the model matrix for random effects.
- Other typical distributions for the conditional random variable $y|u$ are:
  1. The *Bernoulli* distribution for binary data (0 and 1), which density function is:

$$p(y|\mu) = \mu^y (1-\mu)^{1-y}, \quad 0 < \mu < 1, \quad y = 0.1$$

  2. Many independent binary response variables can be represented as a binomial response only if the distributions are Bernoulli and have the same mean
  3. The Poisson distribution for count data (0, 1, ...), which has the density function:

$$p(y|u) = e^{-\mu} \frac{\mu^y}{y!}, \quad 0 < \mu \quad y = 0, 1, 2, \ldots$$

  All these distributions are completely specified by the conditional mean.
- When the conditional distributions $(y|u)$ are constrained by μ, such that $0 < \mu < 1$ (Bernoulli) or $0 < \mu$ (Poisson), we can not define the conditional mean, $\mu_{y|u}$, as equal to a linear predictor, $X\beta + Zu$; since this linear predictor has no restrictions in its resulting values.
- An invertible univariate function g, called **link function**, is selected such that $\eta = g(\mu)$. It is required that g is invertible such that $\mu = g^{-1}(\eta)$ is defined for this $-\inf < \eta < \inf$ and its proper range ($0 < \mu < 1$ for Bernoulli or $0 < \mu for Poisson$).
- The 'link' function for the Bernoulli distribution is the **logit** function.

$$\eta = g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

  The 'link' function for the Poisson distribution is the **log** function.

$$\eta = g(\mu) = \log(\mu)$$

- In the linear mixed model, the parameters estimates are provided through maximizing the likelihood function; i.e. finding estimates that are more likely to fit our data. For generalized linear mixed model, the same procedure is performed. But maximizing the likelihood function is not a simple task since the distributions of our data depend on the mean. To accomplish this maximization **PIRLS** algorithm is implemented.

- Once we estimate the parameters and we have constructed and adequate model for our data, we can compute the fitted values, i.e. model predictions. To find these resulting predictions of the real restricted observations we must implement the inverse function of g to find the real values. For the case of Bernoulli family the inverse function would be

$$\mu = g^{-1}(\eta) = \frac{e^{\eta}}{1+e^{\eta}} = \frac{1}{1+e^{-\eta}}$$

and for the Poisson family

$$\mu = g^{-1}(\eta) = e^{\eta}$$

Analysis using Generalized Linear Mixed Models

The data analyzed is part of the results of a study that was conducted to compare two oral treatments for a certain infection of the big toe nail. The degree of separation of the nail was evaluated in each patient. These patients were randomized into two treatment groups, and seven measurements were performed during seven visits. The first visit at time 0 denotes observation when applying the treatment for the first time. Responses of each of the patients were considered as *0 = no separation*, *1 = separation moderate or severe*, a binary response type. The file containing the data is the **toenail.csv** file. The database contains 104 responses. Not all patients were measured during the seven visits, and therefore some patients have fewer observations than others. They want to take the initial condition of the patient as a random variable.

The codings for treatments are: Treatment A = 0 and Treatment B = 1. This coding is due to the fact that we have a nominal factor and then we have to use **Treatment** contrasts matrix. This encoding is usually used in longitudinal analysis comparing a control group versus a group under a new treatment.

To access the functionality of generalized linear mixed models we have to follow: **Analyze** $->$ **Longitudinal Analysis** $->$ **Generalized Mixed Models**. In Figure 4.40 the main screen is shown.

The model to be fitted is:

$$logit(Y_{it}) = (\beta_0 + \gamma_{02} Treatment_{2i} + b_{0i}) + (\beta_1 + \gamma_{12} Treatment_{1i}) Visit_{ij} + \varepsilon_{it} \tag{4.5}$$

where $Treatment_{2i}$ is a binary variable taking, 1 if patient is taking Treatment B, and 0 it not; β_0 y=and β_1 are, the average intercept and the average slope, respectively, for patient under Treatment A. Treatment A is called of 'reference', since our results will be comparisons against it. γ_{0k} is the average difference of the intercept for Treatment k minus the intercept under Treatment A (reference); γ_{1k} is the average difference of the slopes for patients under Treatment k minus patients under Treatment A; b_{0i} is the random intercept, i.e. patient i initial condition at the beginning of the study. Finally e_{ij} are the residuals of our proposed model. The implementation of this model in LADES is shown in Figure 4.40.

Since our answer is binary (1 or 0's), we have to specify this in the **Options** Section. In Figure 4.41, select the Binomial family to specify our response is binary. Also, select REML as the method for parameter estimation. And finally select the type of contrasts, which in this case is **Treatment**.

Function Outputs

- **Model Matrix.** In this section we can see the model matrix (see A.1) that was used to adjust the model GLMM. This matrix can be seen reflected in the type of contrast that was used, see Figure 4.42.

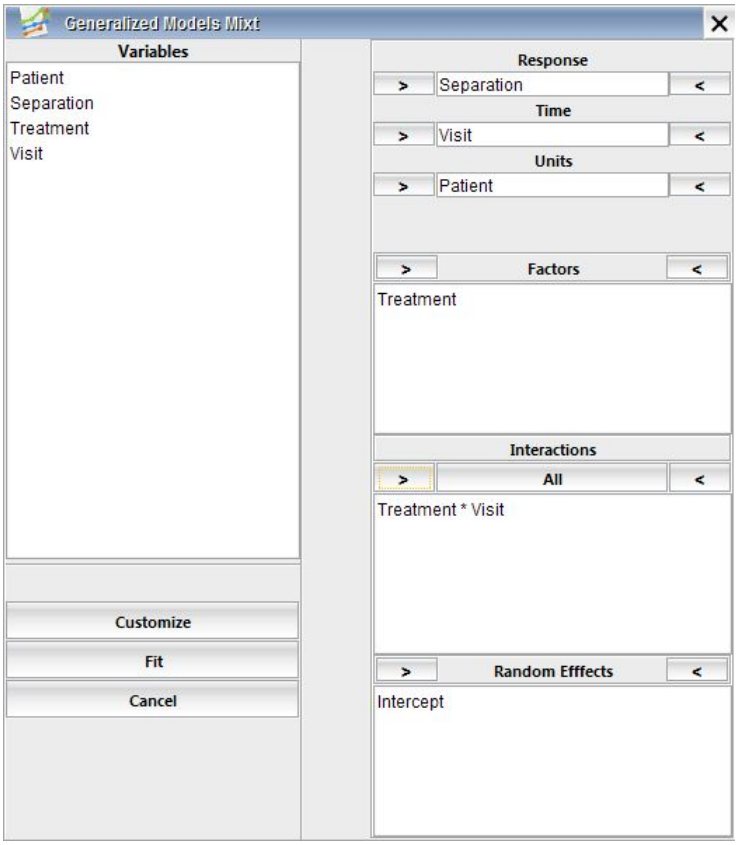


Figure 4.40: Main Screen, GLMM

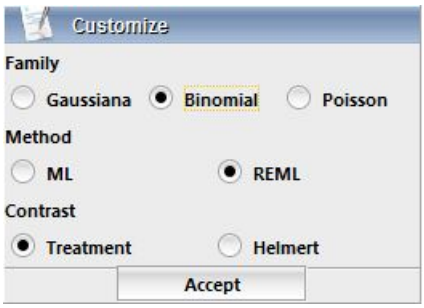
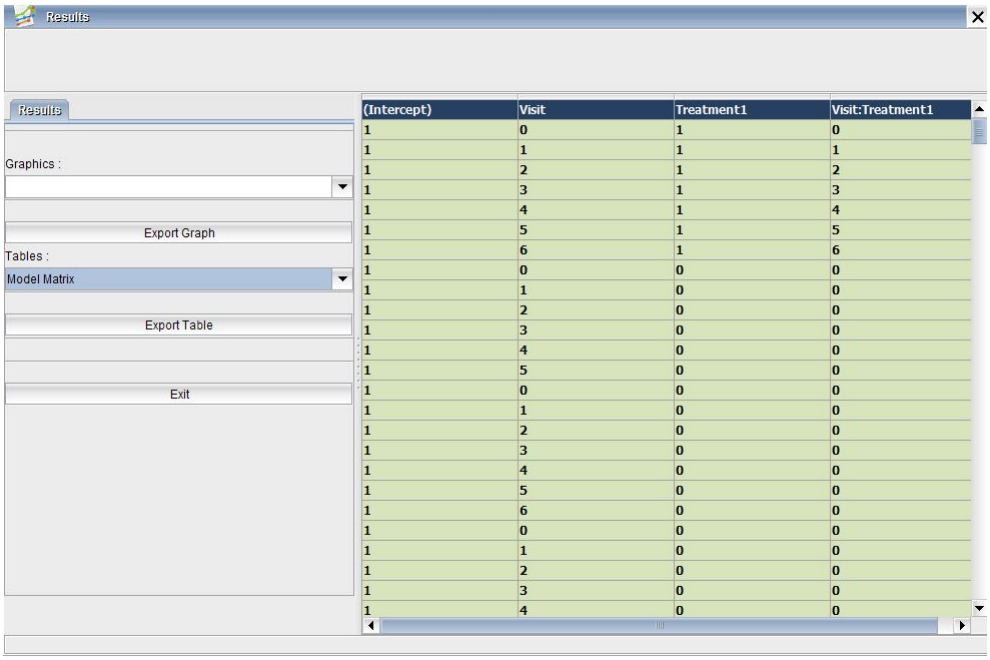


Figure 4.41: Toe Nail data, toenail.csv file



(Intercept)	Visit	Treatment1	Visit:Treatment1
1	0	1	0
1	1	1	1
1	2	1	2
1	3	1	3
1	4	1	4
1	5	1	5
1	6	1	6
1	0	0	0
1	1	0	0
1	2	0	0
1	3	0	0
1	4	0	0
1	5	0	0
1	0	0	0
1	1	0	0
1	2	0	0
1	3	0	0
1	4	0	0
1	5	0	0
1	6	0	0
1	0	0	0
1	1	0	0
1	2	0	0
1	3	0	0
1	4	0	0

Figure 4.42: Model Matrix, GLMM

- **Likelihood Criterion.** In this section the results for the Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) (see A.5) are presented. In addition, the log-likelihood is shown. See Figure 4.43.
- **Fitted Values and Residuals.** In Figure 4.44 adjusted values and residuals for model (4.2.2). The adjusted values are the resulting values after applying the inverse transformation of g. For our case, these values denote the probability to present a separation in the nail of the big toe. The residuals are not restricted, i.e. are based on the 'logit' function.
- **Random Effects Estimation by Subject.** In Figure 4.45, the best linear unbiased predictors for the random coefficients of each patient are displayed. . These predictions were used to calculate the fitted values and residuals. Basically they denote deviations from the overall intercept.
- **Fixed Effects Estimates.** For each of the predictor variables the estimated coefficients, the standard error of the coefficient and the corresponding p-value are given. The p-value is based on the Wald statistic, which is defined as the ratio between the square regression coefficient and its standard error. This statistic follows a χ^2 distribution with 1 degree of freedom. We can observe that **Visit** (Time factor) has an impact of -0.56 per unit time on the logit of the response. And this change is statistically significant (At $\alpha = 0.05$). Due to the type of contrast that was used for **Treatment**, one effect is defined: Treatment B. In our case our profile is Treatment A, as mentioned before. Then, the coefficient of Treatment B refers to the added value given of Treatment B to Treatment A. i.e. the impact of Treatment B has on an patient taking Treatment A. This coefficient is not significant to the 'logit' of the response. The **Treatment:Visit** interaction the adding value in the slope when administrating the Treatment B to a patient taking Treatment A. This interaction produces a change 0.054 in the 'logit' of the response per unit time. We have to mention that our results depend upon the sample size.
- **Parameter estimates by Subject.** Figure 4.47 shows all the estimated coefficient and best linear unbiased predictions for the fixed and random effects, respectively.

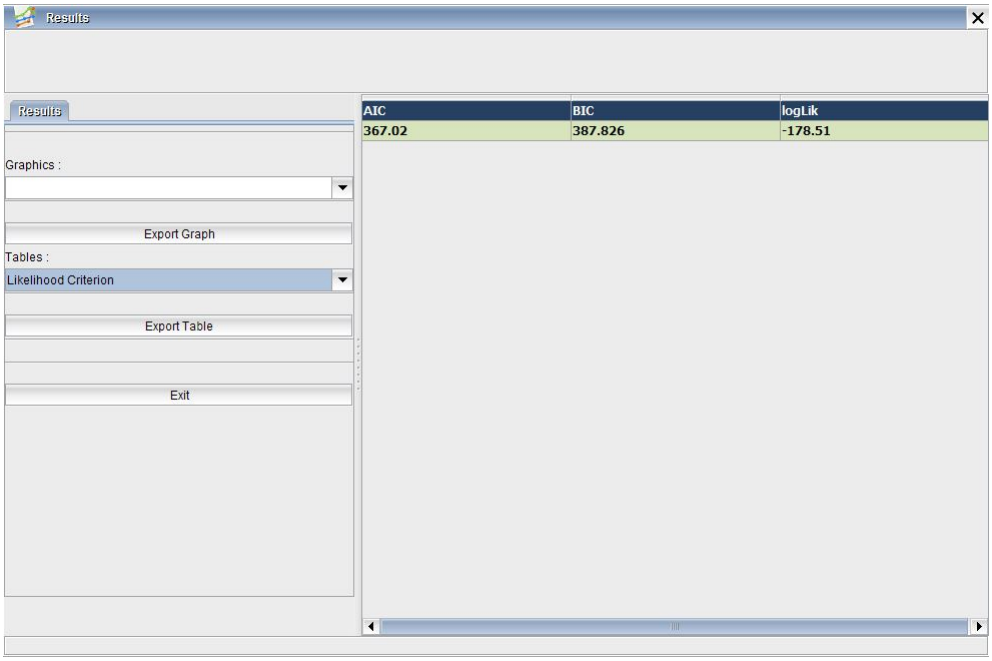


Figure 4.43: Likelihood Criteria, GLMM

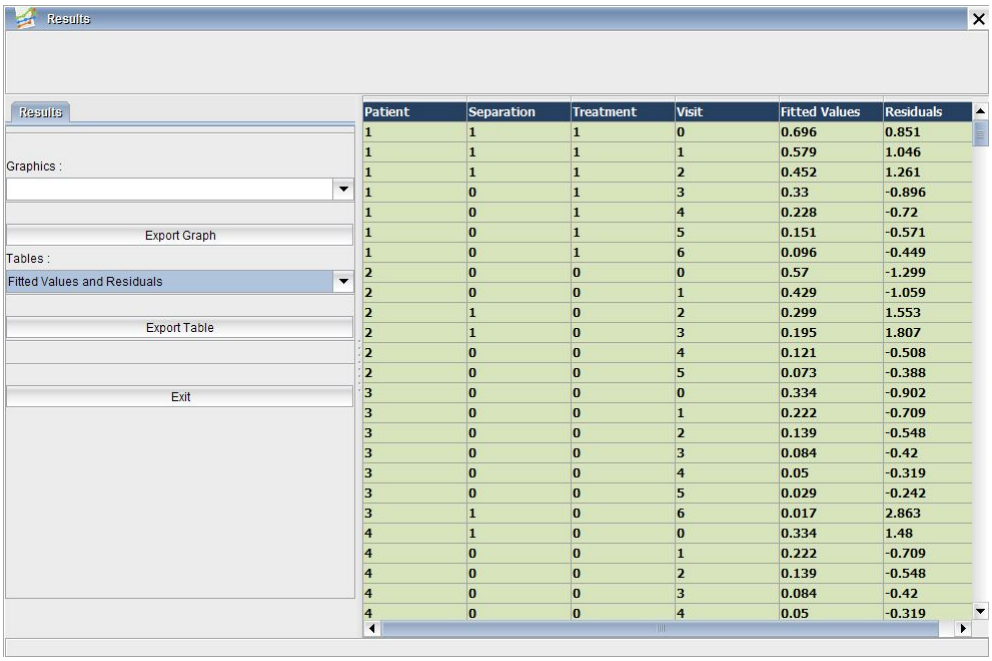
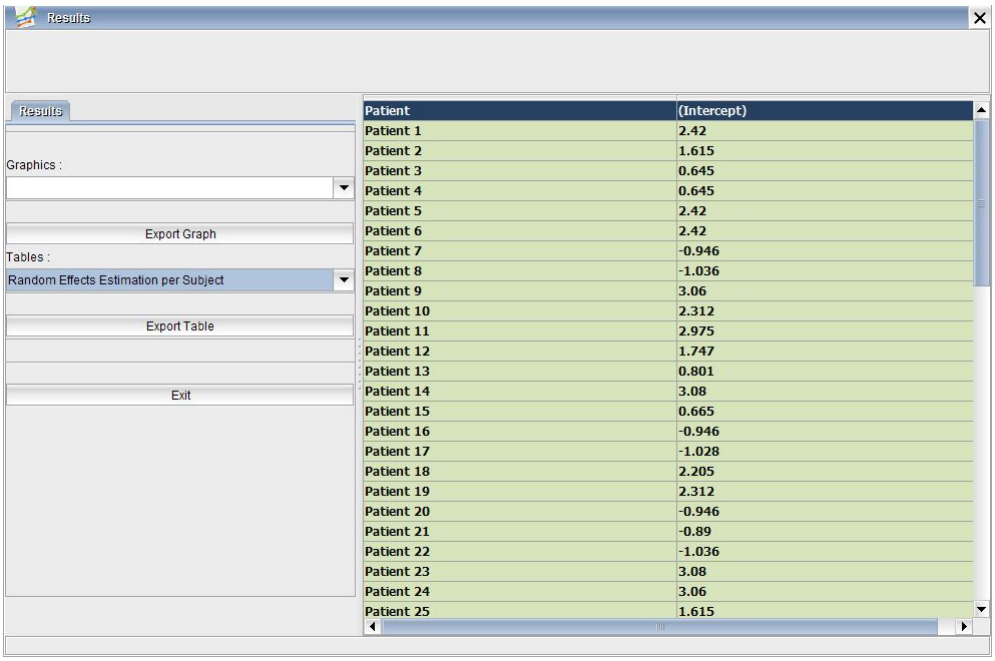
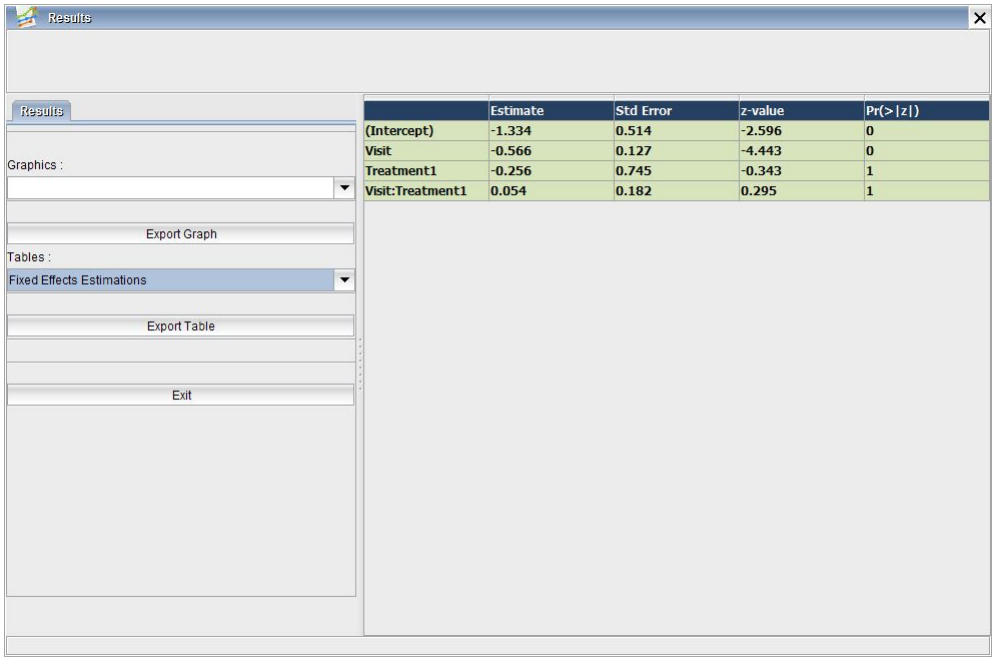


Figure 4.44: Adjusted Values and Residuals, GLMM



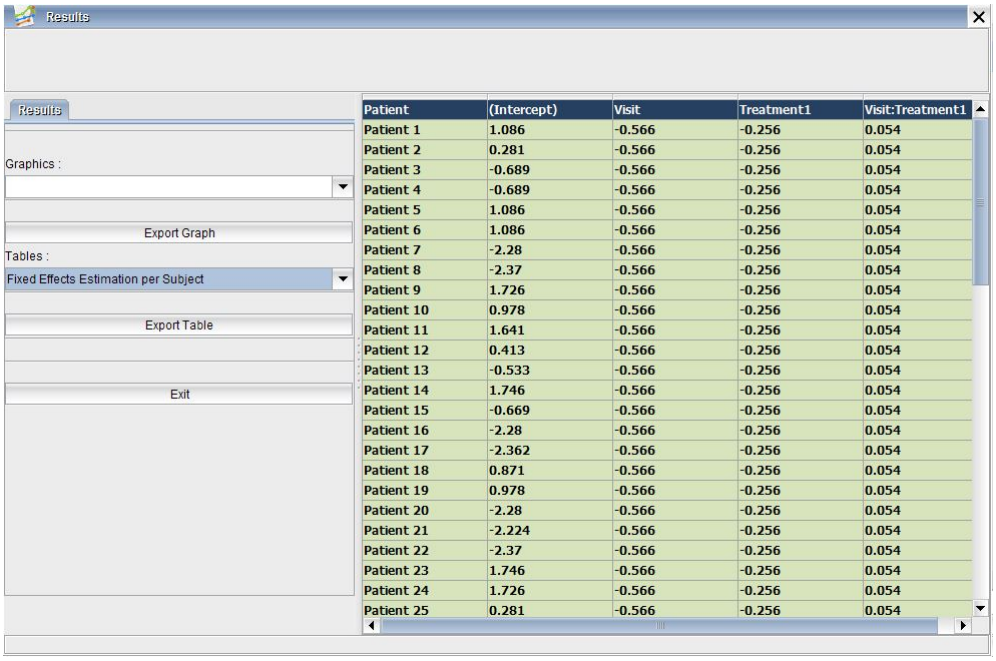
Patient	(Intercept)
Patient 1	2.42
Patient 2	1.615
Patient 3	0.645
Patient 4	0.645
Patient 5	2.42
Patient 6	2.42
Patient 7	-0.946
Patient 8	-1.036
Patient 9	3.06
Patient 10	2.312
Patient 11	2.975
Patient 12	1.747
Patient 13	0.801
Patient 14	3.08
Patient 15	0.665
Patient 16	-0.946
Patient 17	-1.028
Patient 18	2.205
Patient 19	2.312
Patient 20	-0.946
Patient 21	-0.89
Patient 22	-1.036
Patient 23	3.08
Patient 24	3.06
Patient 25	1.615

Figure 4.45: Random Effects Estimates by Subject, GLMM



	Estimate	Std Error	z-value	Pr(>|z|)
(Intercept)	-1.334	0.514	-2.596	0
Visit	-0.566	0.127	-4.443	0
Treatment1	-0.256	0.745	-0.343	1
Visit:Treatment1	0.054	0.182	0.295	1

Figure 4.46: Fixed Effects Estimates, GLMM



Patient	(Intercept)	Visit	Treatment1	Visit:Treatment1
Patient 1	1.086	-0.566	-0.256	0.054
Patient 2	0.281	-0.566	-0.256	0.054
Patient 3	-0.689	-0.566	-0.256	0.054
Patient 4	-0.689	-0.566	-0.256	0.054
Patient 5	1.086	-0.566	-0.256	0.054
Patient 6	1.086	-0.566	-0.256	0.054
Patient 7	-2.28	-0.566	-0.256	0.054
Patient 8	-2.37	-0.566	-0.256	0.054
Patient 9	1.726	-0.566	-0.256	0.054
Patient 10	0.978	-0.566	-0.256	0.054
Patient 11	1.641	-0.566	-0.256	0.054
Patient 12	0.413	-0.566	-0.256	0.054
Patient 13	-0.533	-0.566	-0.256	0.054
Patient 14	1.746	-0.566	-0.256	0.054
Patient 15	-0.669	-0.566	-0.256	0.054
Patient 16	-2.28	-0.566	-0.256	0.054
Patient 17	-2.362	-0.566	-0.256	0.054
Patient 18	0.871	-0.566	-0.256	0.054
Patient 19	0.978	-0.566	-0.256	0.054
Patient 20	-2.28	-0.566	-0.256	0.054
Patient 21	-2.224	-0.566	-0.256	0.054
Patient 22	-2.37	-0.566	-0.256	0.054
Patient 23	1.746	-0.566	-0.256	0.054
Patient 24	1.726	-0.566	-0.256	0.054
Patient 25	0.281	-0.566	-0.256	0.054

Figure 4.47: Parameter estimates by Subject, GLMM

- **Estimation of Random Effects Parameter.** In this section the estimation of the random effects parameter is shown. Only one parameter was estimated, σ_{b_0}, since the distribution of the random intercept is Normal with mean zero. Figure 4.48 shows such estimation $\hat{\sigma_{b_0}}$.
- **Correlation Fixed Effects.** Shows the correlation of the fixed effects used in our model. See Figure 4.49.

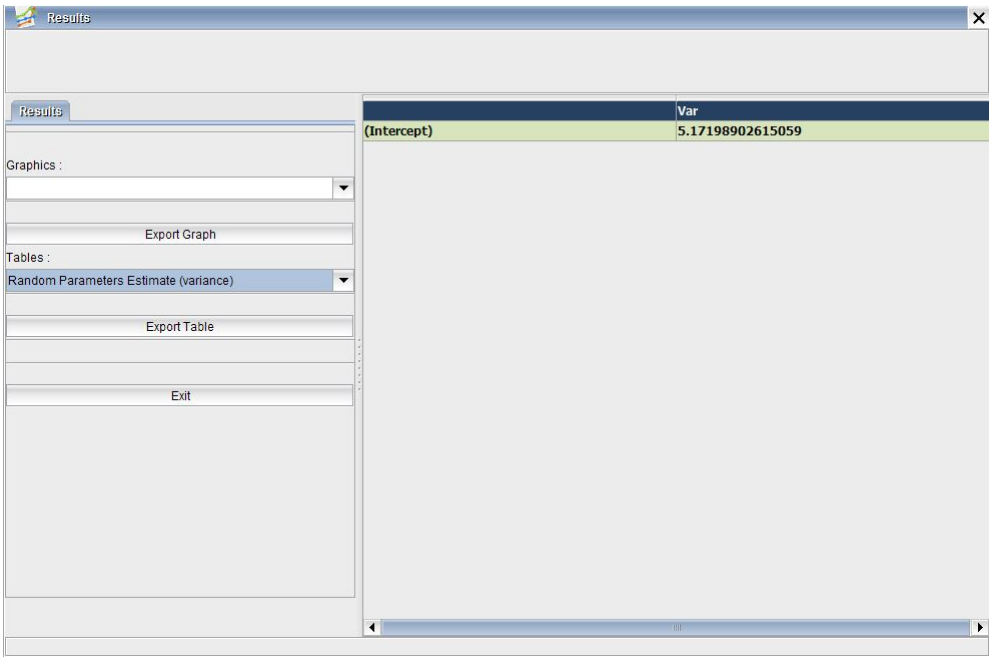


Figure 4.48: Estimation of Random Effects Parameter (Variance), GLMM

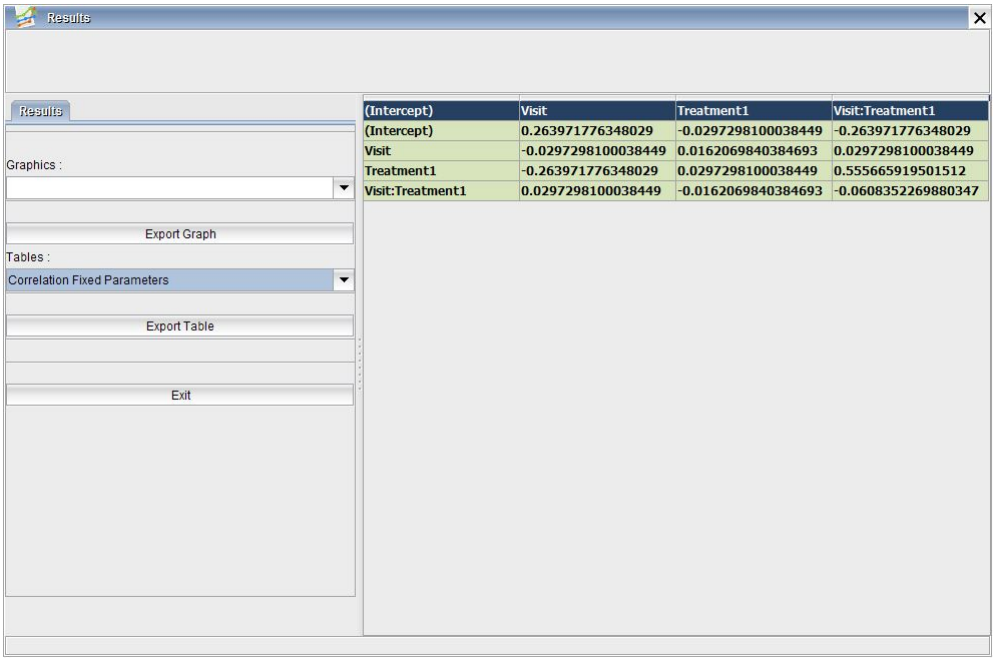


Figure 4.49: Correlation of Fixed Parameters, GLMM

4.2.3 Generalized Estimating Equations

The method of Generalized Estimating Equations (GEE) allows us to model variables' relations at different points in time in a simultaneous way. Then, the estimate of β_1 in equation 4.2.3, is reflecting the relationship between the longitudinal development in time of the response variable Y and the longitudinal development of the corresponding predictor variable X_1.

$$Y = \beta_0 + \beta_1 X + \ldots + \varepsilon \tag{4.6}$$

GEE is an iterative procedure which uses the quasi-likelihood to estimate regression coefficients.

Summary of Generalized Estimating Equations

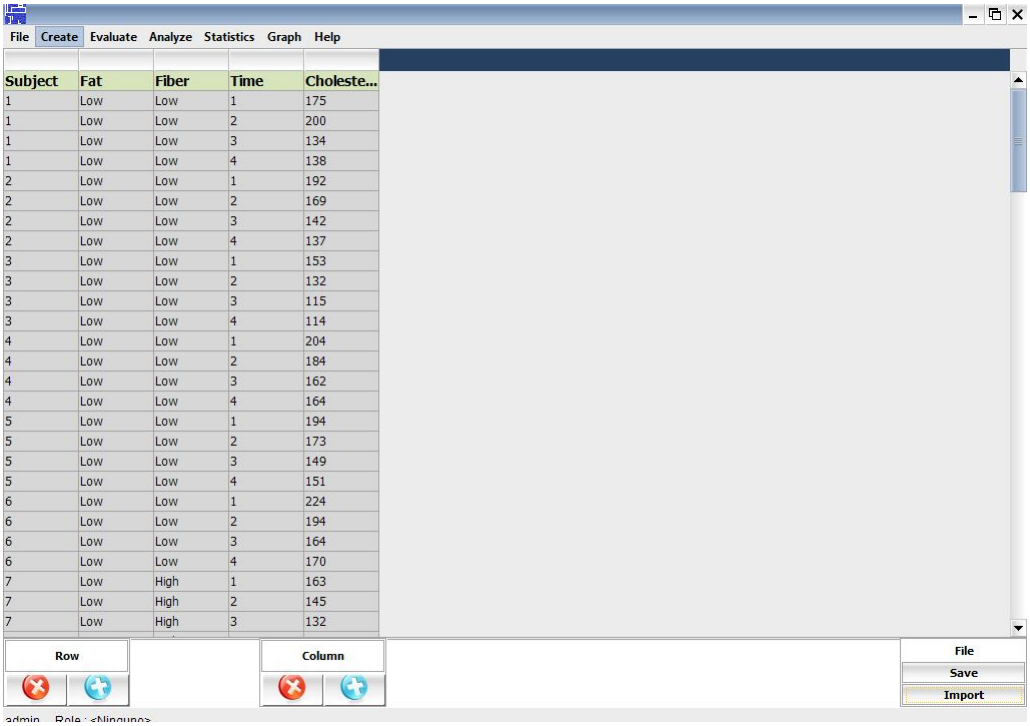
- The literature assumes that GEE analysis is robust against wrong-selection of the correlation matrix structure. However, we must be careful when selecting the best option for our correlation matrix (see A.4). Unfortunately, there is not a direct method to determine what correlation structure is more appropriate. One possibility is to analyze the correlation structure of the within-subject (observations over time) observed data; and then, find the most appropriate structure.
- The simplicity of the correlation structure is a factor we must take into account. The number of parameters (in this case, correlation coefficients) that need to be estimated is different according to the structure we choose. For instance, in exchangeable structure we estimate only one parameter for all the observations in time; while for a complete structure with, for example, five observations over time, we estimate five parameters. Finally, the power of statistical analysis will be influenced by the choice of the structure.
- The parameter estimation process can be seen as follows. First, simple linear regression a model assuming that the observations are independent is adjusted. Then, based on this residual analysis, the parameters for the correlation matrix are computed. Finally, the regression coefficients are re-estimated, using this correlation matrix to correct residuals dependency. The process is repeated until it converges.
- In GEE, the correlation within-subjects is considered a "nuisance" parameter. Then, GEE corrects the with-in-subjects dependency by equation 4.2.3.

$$Y_{it} = \beta_0 + \beta_1 X_{it\,j} + \beta_2 t + \cdots + CORR_{it} + \varepsilon_{it} \tag{4.7}$$

where Y_{it} is the observations of subject i at time t, β_0 is the intercept, $X_{i\,jt}$ is the independent variable j for the individual i at time t, $\beta_{1\,j}$ is the regression coefficient for independent variable j, J is the number of independent variables, t is time, β_2 is the regression coefficient for time, $CORR_{it}$ represents the elements of the correlation structure, and ε_{it} is the error for subject i at time t.

Longitudinal Análysis using GEE

A study was conducted to evaluate the effect of six diets in cholesterol levels in older adults. Diets were constructed from six combinations of two fiber levels (Low, High) and three levels of fat (Low, Medium, High). Thirty-six people were randomly assigned to these six diets. Each level of cholesterol was determined every two months (Time 1 = 2 months, Time 2 = 4 Months, Time 3 = 6 months time 4 = 8 months). To access the GEE function in LADES we follow



Subject	Fat	Fiber	Time	Choleste...
1	Low	Low	1	175
1	Low	Low	2	200
1	Low	Low	3	134
1	Low	Low	4	138
2	Low	Low	1	192
2	Low	Low	2	169
2	Low	Low	3	142
2	Low	Low	4	137
3	Low	Low	1	153
3	Low	Low	2	132
3	Low	Low	3	115
3	Low	Low	4	114
4	Low	Low	1	204
4	Low	Low	2	184
4	Low	Low	3	162
4	Low	Low	4	164
5	Low	Low	1	194
5	Low	Low	2	173
5	Low	Low	3	149
5	Low	Low	4	151
6	Low	Low	1	224
6	Low	Low	2	194
6	Low	Low	3	164
6	Low	Low	4	170
7	Low	High	1	163
7	Low	High	2	145
7	Low	High	3	132

Figure 4.50: colesterol.csv dataset

Analyze $->$ **Longitudinal Analysis** $->$ **GEE**. In Figure 4.50 we observe **cholesterol.csv** data already loaded into LADES.

Now, Figure 4.51 shows the screen where we have to input the model specifications. Fields such as **Response**, **Time** and **Subject** are mandatory. Fields such as **Factors** and **Interactions** define a more elaborated model. On the other hand, Figure 4.52 displays the customization section to especify other parameters like: **Family** (response variable), **Correlation Structure** (correlation between observations over time), and **Contrasts** (which have an impact on the interpretation of the estimated coefficients)

For our problem, we opted to fit the model

$$Y_{it} = \beta_0 + \beta_1 t + \beta_3^* Fat + \beta_4^* Fiber + \beta_5^* Fat * Time + \beta_6^* Fiber * Time + CORR_{it} + \varepsilon_{it} \quad (4.8)$$

where β^* means this coefficient is splitter to several effects, since we choose *Treatment* as our type of contrast (see A.2). This will be explained later. We also chose well (see Figure 4.52) an Exchangeable correlation structure since this requires fewer parameters to estimate, and was suggested by the researchers conducting the study. Finally, a Gaussian family was selected.

Function Outputs

The resulting tables are:

- **Hypothesis Testing for Parameters.** For each of the predictor variables the regression coefficient, the standard error of the coefficient and corresponding p-value are given. The p-value is based on the Wald statistic, which is defined as the ratio between the square regression coefficient and its standard error. This statistic follows a χ^2 distribution with 1 degree of freedom. We can observed that **Time** has a negative impact of cholesterol of

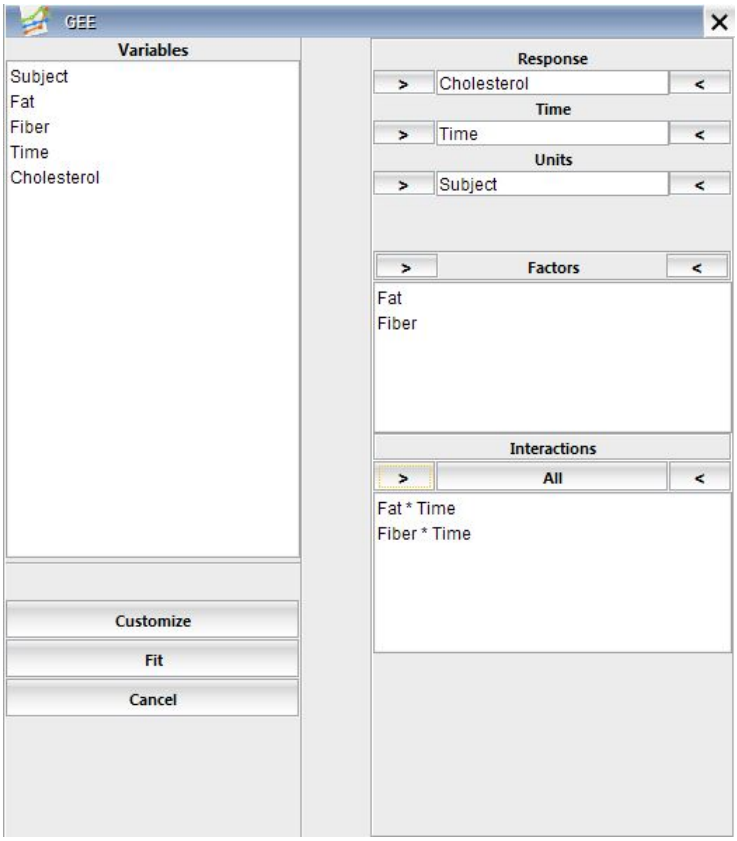
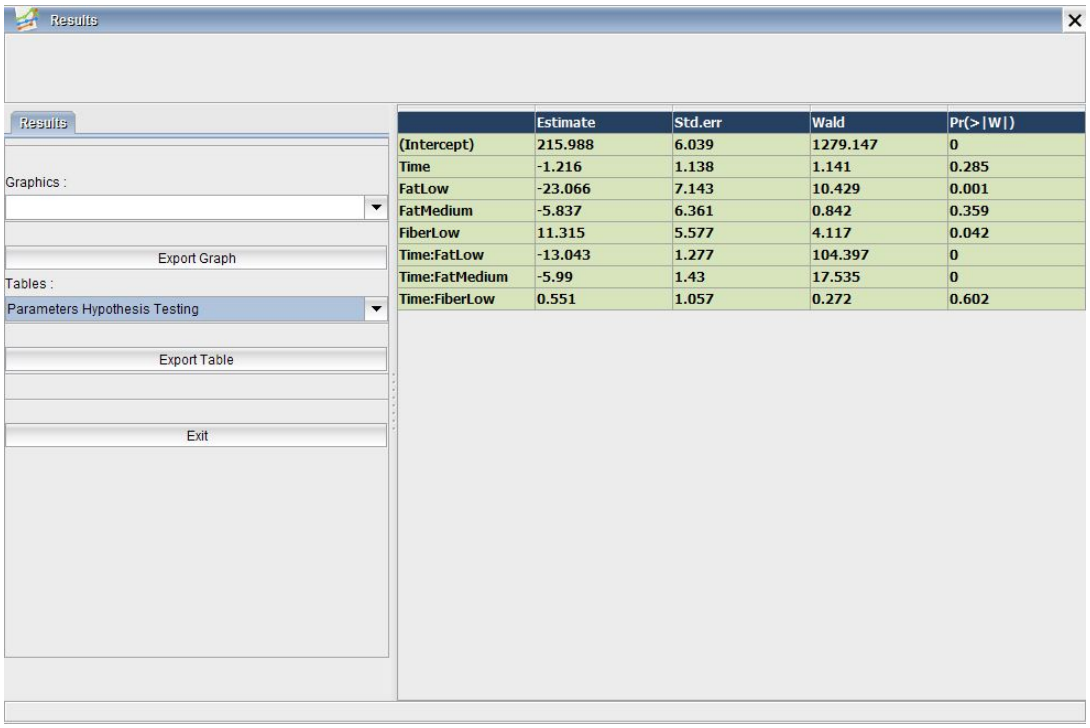


Figure 4.51: Main Screen, GEE

Figure 4.52: Customization Screen, GEE

1,216 per units each unit time. But this change is not statistically significant at a level of α = 0.05. Since the type of contrast that was used for this model, **Fat** factor was splitted into two effects: LowFat and MediumFat. In this case our profile or reference level for Fat is the High level. As a result, the coefficient for LowFat refers to the change that occurs when following a diet high in fats and add the effect of Low Fat diet. This coefficient turns out to be significant and its estimated value is -23.06 units of cholesterol, that is, the impact of a low fat diet on a high fat diet is negative. In contrast, for MediumFat coefficient, the change that occurs is about -5.837 and it is not statistically significant. This indicates that there is no difference if you decide to change from a high fat diet to a medium fat diet . For **Fiber**, the reference level is High. Then the coefficient for LowFiber factor refers to the effect produced pf being on a diet with high content of fiber and add a diet with content. This change is of 11,315 units in colesterol, but this is not significant. Therefore, there is no difference if we choose a High Fiber Diet or a Low fiber diet. The interaction factor **Time:LowFat** refers to the change in the growth over time produced by adding a low-fat diet to a high fat. This interaction is not significant. The rest of the interactions are not significant.



	Estimate	Std.err	Wald	Pr(>|W|)
(Intercept)	215.988	6.039	1279.147	0
Time	-1.216	1.138	1.141	0.285
FatLow	-23.066	7.143	10.429	0.001
FatMedium	-5.837	6.361	0.842	0.359
FiberLow	11.315	5.577	4.117	0.042
Time:FatLow	-13.043	1.277	104.397	0
Time:FatMedium	-5.99	1.43	17.535	0
Time:FiberLow	0.551	1.057	0.272	0.602

Figure 4.53: Hypothesis Testing for Parameters, GEE

- **Model Matrix.** In this section we can see the model matrix (see A.1) that was used to adjust the GEE model. We can see the contrasts we used, see Figure 4.54.
- **Scale Parameter Estimation.** Refers to the relationship between the variance of the observations and their expected value. This parameter estimation is shown in Figure 4.55. Because we are not making assumptions about the distribution of our observations, this parameter is simply capturing the information concerning the variability of our observations with respect to its expected value
- **Correlation Parameter Estimations.** Because it was decided to use an 'order 1 autoregressive' (**AR1**) correlation structure in our data, only one parameter was estimated. Figure

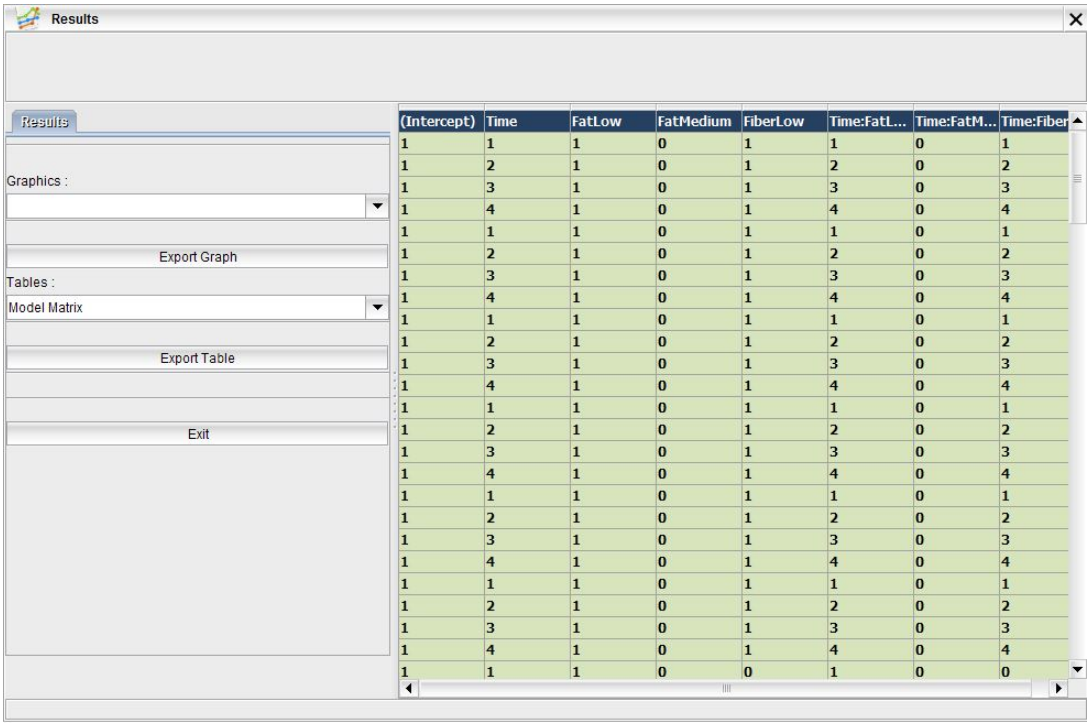


Figure 4.54: Model Matrix, GEE

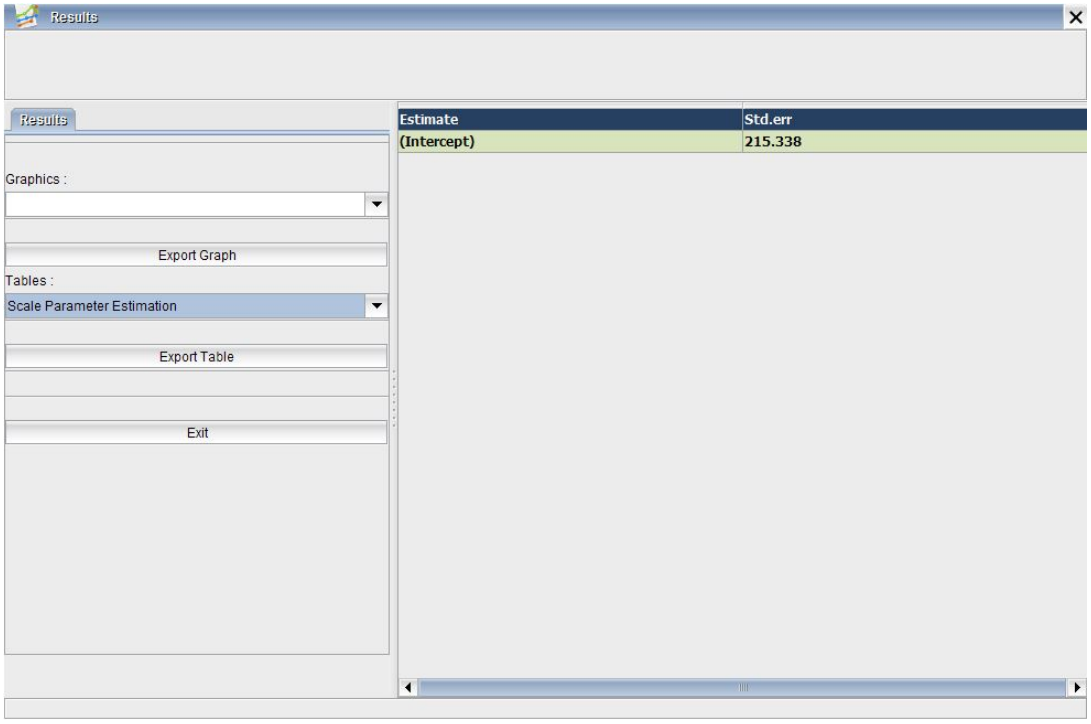


Figure 4.55: Scale Parameter Estimation, GEE

4.56 shows the result of this estimate. The value of the parameter correlation is 0.841. This means adjacent observations, i.e. whose distance is equal to 1, have a correlation equal to 0.841. On the other hand, observations with two units of distance, for example the observation at time 1 and the observation time 3, have a correlation of $0.841^2 = 0.707$. And the observations whose separation is 3 units time have correlaciion of $0,841^3 = 0.594$. This means say that the correlation decreases as more distant observations are.

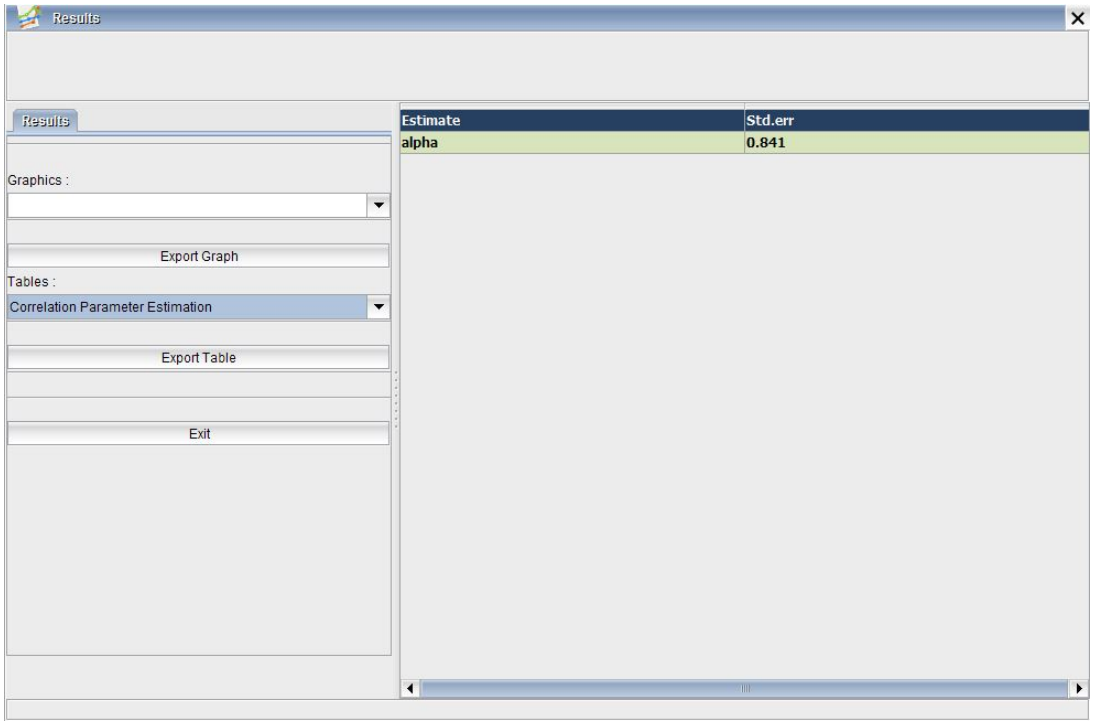
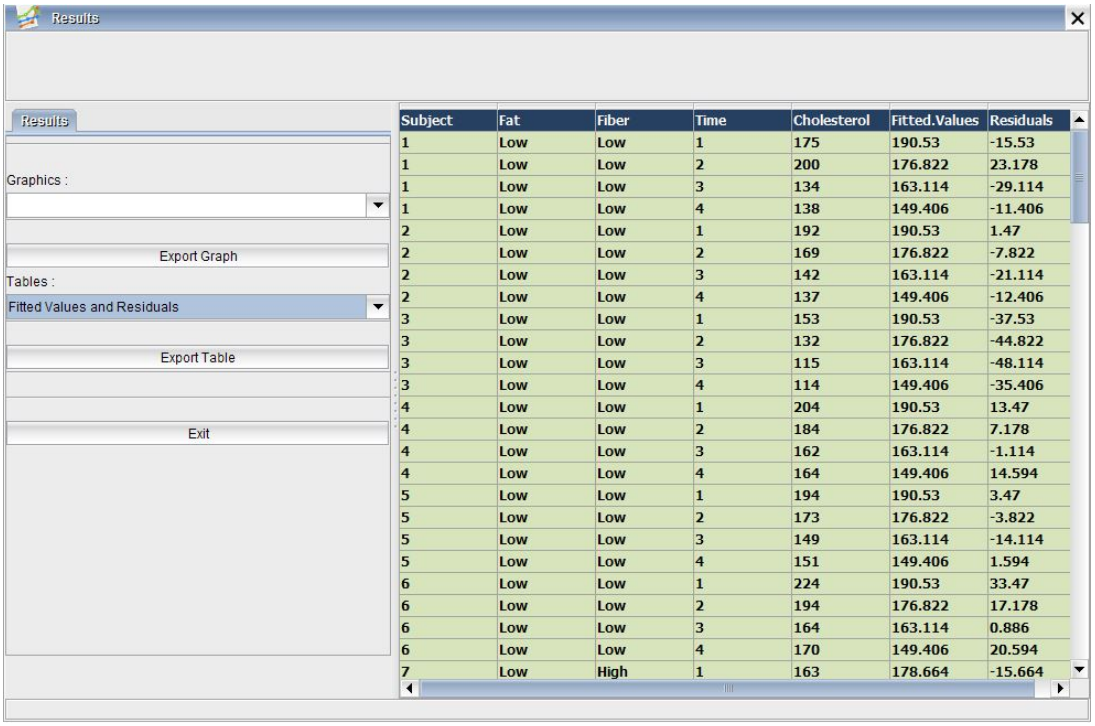


Figure 4.56: Correlation Parameter Estimations, GEE

- **Fitted Values and Residuals.** Finally, we show the adjusted values and residuals in Figure4.57.



Subject	Fat	Fiber	Time	Cholesterol	Fitted.Values	Residuals
1	Low	Low	1	175	190.53	-15.53
1	Low	Low	2	200	176.822	23.178
1	Low	Low	3	134	163.114	-29.114
1	Low	Low	4	138	149.406	-11.406
2	Low	Low	1	192	190.53	1.47
2	Low	Low	2	169	176.822	-7.822
2	Low	Low	3	142	163.114	-21.114
2	Low	Low	4	137	149.406	-12.406
3	Low	Low	1	153	190.53	-37.53
3	Low	Low	2	132	176.822	-44.822
3	Low	Low	3	115	163.114	-48.114
3	Low	Low	4	114	149.406	-35.406
4	Low	Low	1	204	190.53	13.47
4	Low	Low	2	184	176.822	7.178
4	Low	Low	3	162	163.114	-1.114
4	Low	Low	4	164	149.406	14.594
5	Low	Low	1	194	190.53	3.47
5	Low	Low	2	173	176.822	-3.822
5	Low	Low	3	149	163.114	-14.114
5	Low	Low	4	151	149.406	1.594
6	Low	Low	1	224	190.53	33.47
6	Low	Low	2	194	176.822	17.178
6	Low	Low	3	164	163.114	0.886
6	Low	Low	4	170	149.406	20.594
7	Low	High	1	163	178.664	-15.664

Figure 4.57: Valores Ajustados y Residuos, GEE

4.2.4 Generalized Least Squares, GLS

In linear mixed models, the covariance matrix of the response vector y_i is

$$Var(y_i) = \sum_i = \sigma^2 \left(Z_i D Z_i^T + \Lambda_i\right)$$

this matrix has two components which can be used to model *heteroskedasticity* and *correlation*: a component of random effects, given by $Z_i D Z_i^T$, and a component within-subjects, given by Λ_i, respectively.

In some applications it is desirable to avoid the incorporation of random effects in the model to reflect the dependency between observations. If we only use the within-subjects component, Λ_i, to model the structure of the response variance, we can generate a simplified version of a linear random effects model. That is

$$y_i = X_i\beta + \varepsilon_i, \quad \varepsilon \sim N(0, \sigma^2\Lambda_i), \quad i = 1, \ldots, M. \tag{4.9}$$

The parameter estimation under this model has been studied in the linear regression literature. It is usually assumed that Λ_i is known. This problem of estimation is known as the generalized least squares problem.

Summary of GLS

- If we use the transformation

$$y_i^* = \left(\Lambda_i^{-1/2}\right)^T y_i; \quad X_i^* = \left(\Lambda_i^{-1/2}\right)^T X_i; \quad \varepsilon_i^* = \left(\Lambda_i^{-1/2}\right)^T \varepsilon_i$$

where Λ_i is positive definite and $\Lambda_i^{-1} = \Lambda_i^{-1/2}\left(\Lambda_i^{-1/2}\right)^T$. We can re-express the model in Equation 4.2.4 as a classical linear regression model

$$y_i^* = X_i^*\beta + \varepsilon_i^*, \quad \varepsilon_i^* \sim N(0, \sigma^2 I), \quad i = 1, \ldots, M \tag{4.10}$$

Then, for a fixed λ, the estimates of the parameters β and σ^2 can be obtained by ordinary least squares (OLS).

- The parameter estimation is performed as follows. First we have to find least squares estimators for the parameters β and σ^2 for model (4.2.4); these estimations are conditioned to the value of λ. Then, the function log-likelihood of the model (4.2.4) is constructed by substituting the estimates $\beta(\lambda)$and $\sigma^2(\lambda)$. Now, the likelihood function is a function only of λ. Finally, the likelihood function is maximized over λ, i.e. the maximum likelihood estimator $\hat{\lambda}$ is founded and the conditioned estimators are $\beta(\hat{\lambda})$ and $\sigma^2(\hat{\lambda})$ obtained. We can restrict the likelihood function for model (4.2.4). This restriction consists on integrating the β vector out from the full likelihood function. Then, after de maximization, the resulting estimators $\hat{\beta}$ and $\hat{\sigma}^2$, can be obtained. Such estimators are known as restricted maximum likelihood estimators.
- The Λ_i matrices can be decomposed into a set of simple matrices: $\Lambda_i = V_i C_i V_i$. Where V_i describes the variance structure and C_i describes the correlation structure of the within-subject errors ε_i. That is, the variance and the correlation matrix of each individual i.

Analysis using GLS

As an example we have an experiment where the distance from the pituitary to the pterygo-maxilar fisure (mm) was studied. The distances were obtained from a X-ray to each of the skulls of individuals. Men and women were studied at ages 8, 10, 12 and 14 years. To avoid problems with the estimates and the interpretations of the results, it was decided to codify the age variable as centered around 11 years. Then, the analysis is aimed to evaluate the growth of such distance around the age of 11 years. The file containing the results of the study is **orthodoncy.csv**. They want to set the following model

$$distance_{ij} = \beta_0 + \beta_1 Sex + \beta_2(Age_j - 11) + \beta_3(Age_j - 11) : Sex + \varepsilon_{ij} \tag{4.11}$$

where $i = 1, \ldots, 27$ individuals, and measured points $j = 1, \ldots, 4$ (-3, -1, 1, 3).

To access the generalized least squares function just follow **Analyze** $->$ **Longitudinal Analysis** $->$ **GLS**. The main screen of this function is shown in Figure 4.58.

In Figure 4.58 we can see the model (4.2.4) implemented in LADES. It was decided to use a **Exchangeable** correlation structure, i.e. the correlation between observations over time for each of the individuals is the same (see A.4). We select this in the **Options** section, see Figure 4.59. Moreover we used **Helmert** contrasts (see A.2), since the age variable contains negative and positive values. Such contrasts encode **sex** as *male = -1* and *female = 1* so that our model matrix is orthogonal.

We will model this phenomenon using the Ident Variance Structure using **Sex** (see A.3) and a General correlation structure.

After setting this new model we proceed to the Function Outputs.

Function Outputs

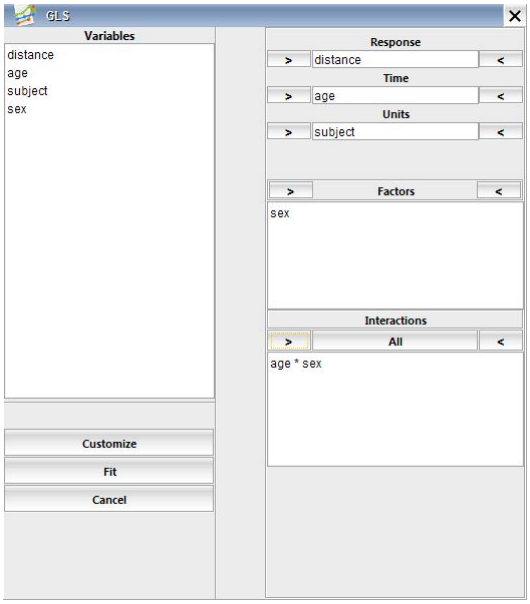


Figure 4.58: Main Screen, GLS

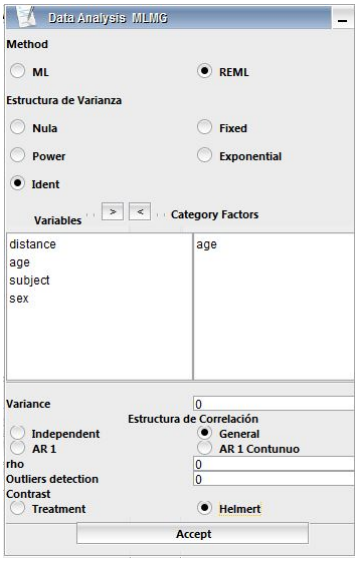


Figure 4.59: Options Screen, GLS

- **95 % Confidence Intervals for Fixed Parameter Estimates.** Figure 4.60 shows 95% confidence intervals and estimates for the parameter in model (4.2.4). In addition, confidence intervals for correlation and variance structure parameters are shown. A first evaluation to determine which factors are significant is that its confidence interval does not include zero

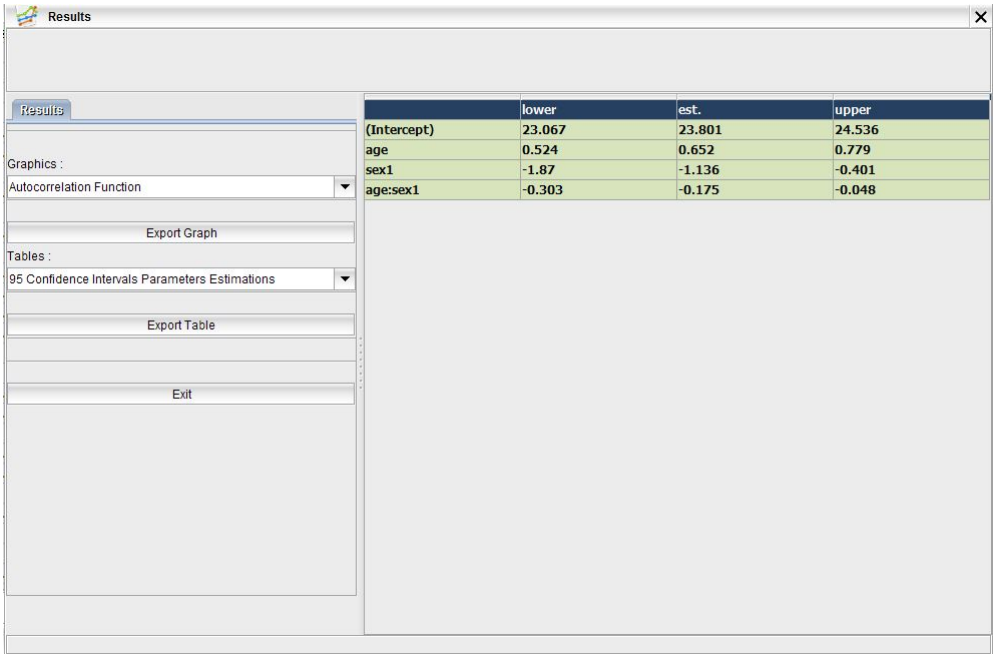


Figure 4.60: Confidence Intervals of Parameter Estimates, GLS

- **95 % Confidence Intervals for Correlation Parameter Estimates.** Figure 4.61 shows 95% confidence intervals and estimates for each parameter in the general correlation structure. For instance, cor(1:2) represents the estimate of parameter $\rho_{(1,2)}$ in the correlation structure.
- **95 % Confidence Intervals for Variance Parameter Estimates.** Figure 4.62 shows 95% confidence intervals and estimates for parameters in the variance structure. For example, -1 represents the estimation of the variance for the first stratification level.
- **95 % Confidence Intervals for Variance Residuals Estimate.** Figure 4.63 shows 95% confidence interval and estimate for the variance of the error.
- **Likelihood Criteria.** In this section the resulting values for the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) (see A.5) are presented. The log likelihood is also shown. See Figure 4.64.
- **Residuals Statistics.** In this section, see Figure 4.65, shows some statistics of the model residuals. We note that the median of the residuals is close to 0, and Quartile 1 and Quartile 4 to the minimum and maximum, respectively. This gives evidence that the residuals may follow a symmetric distribution around zero.
- **Fixed Effects Correlations.** In Figure 4.66, we observe the correlation matrix of the estimates of the fixed coefficients.
- **Adjusted Values and Residuals.** In Figure 4.67 adjusted values and residuals of model (4.2.4) are shown.
- **Fixed Effects and Hypothesis Testing.** In Figure 4.68 the estimates of the coefficient, the standard error and the corresponding p-value are presented. The p-value is based in

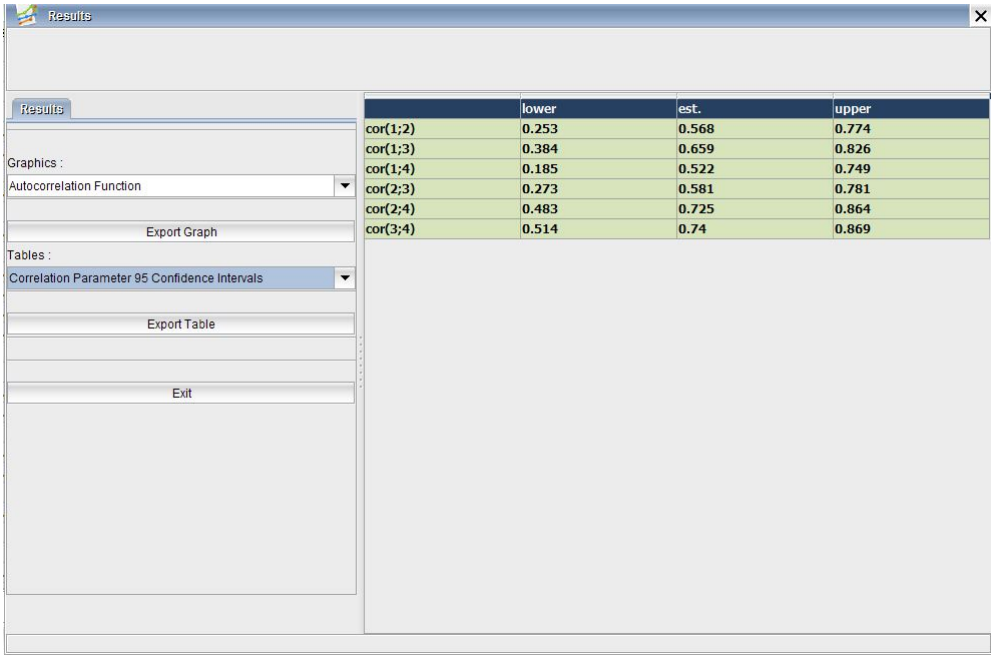


Figure 4.61: Confidence Intervals for Correlation Parameter Estimates, GLS

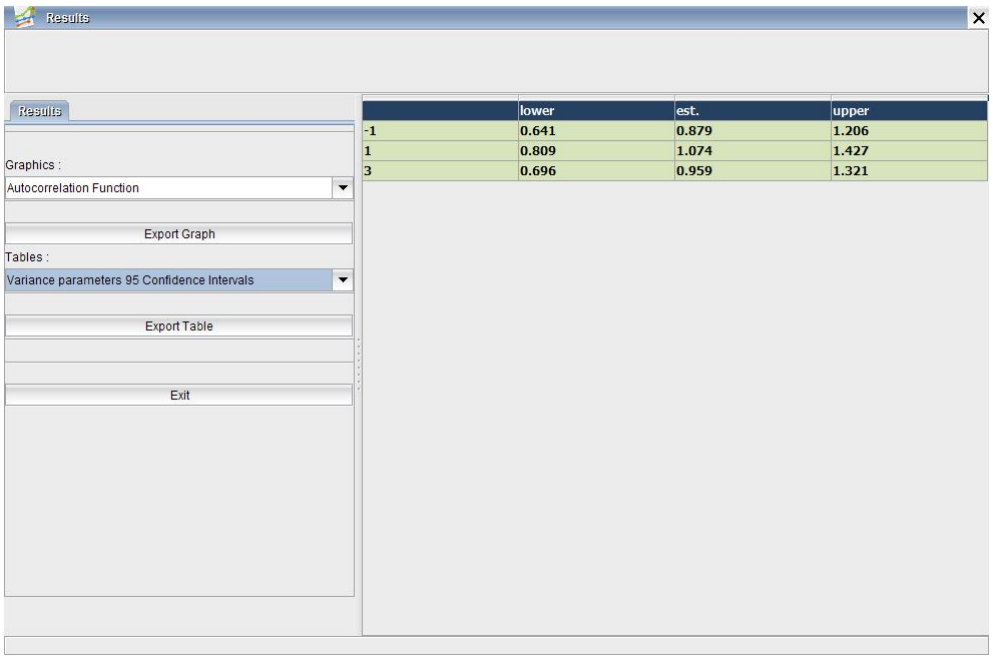


Figure 4.62: Confidence Intervals for Variance Parameter Estimates, GLS

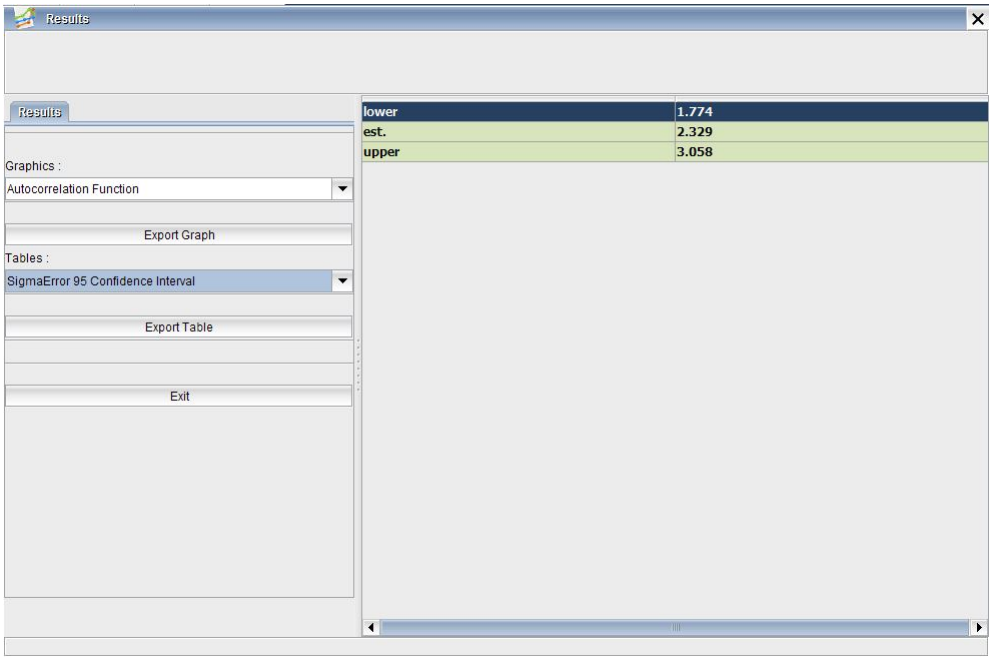


Figure 4.63: Confidence Intervals for Variance Error Estimate, GLS

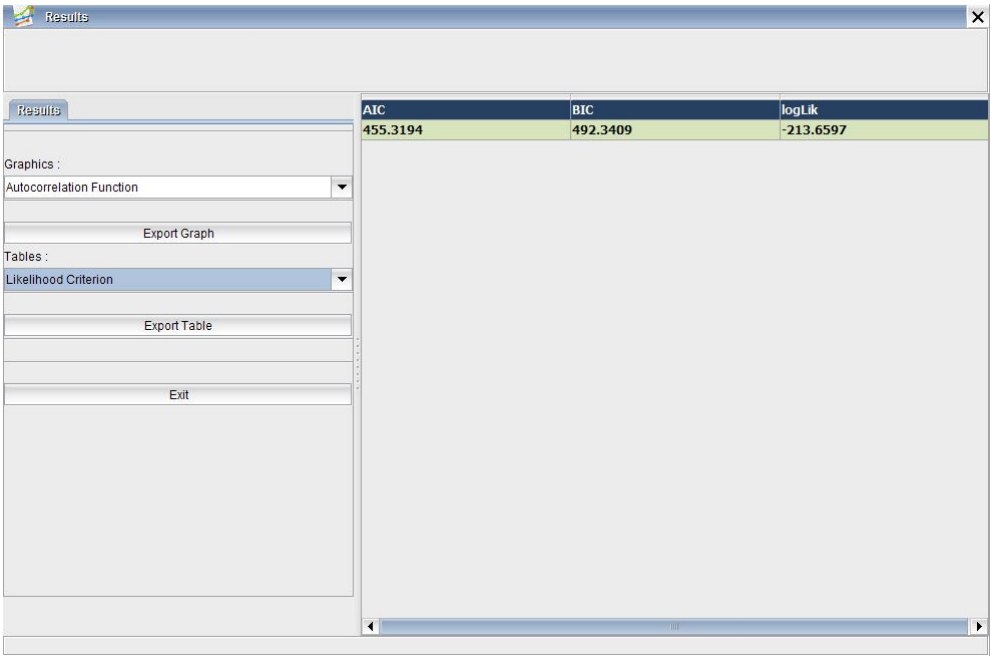


Figure 4.64: Likelihood Criteria, GLS

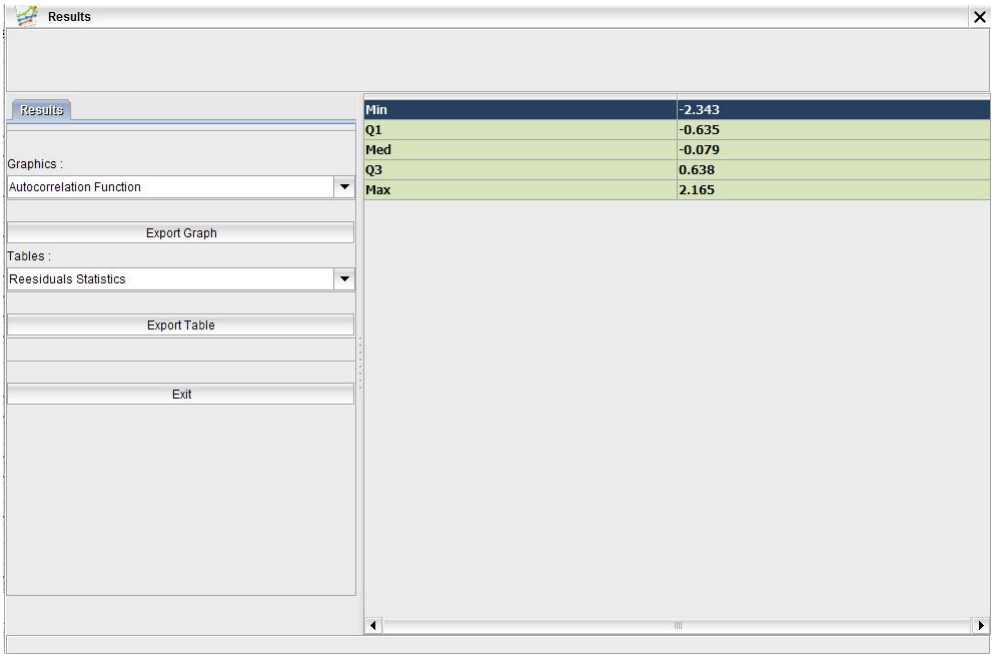


Figure 4.65: Residuals Statistics, GLS

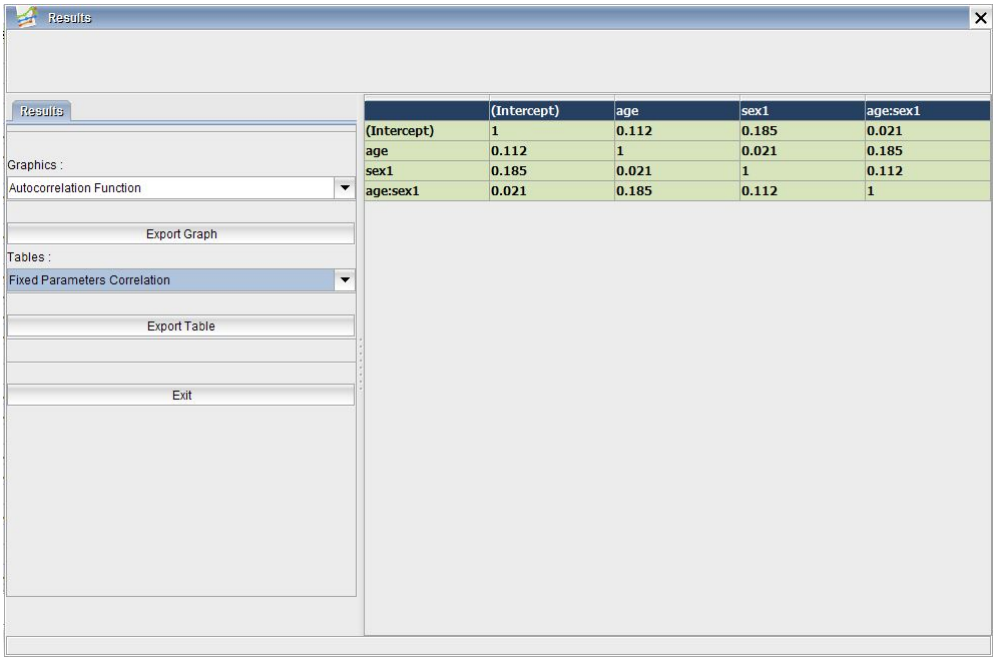
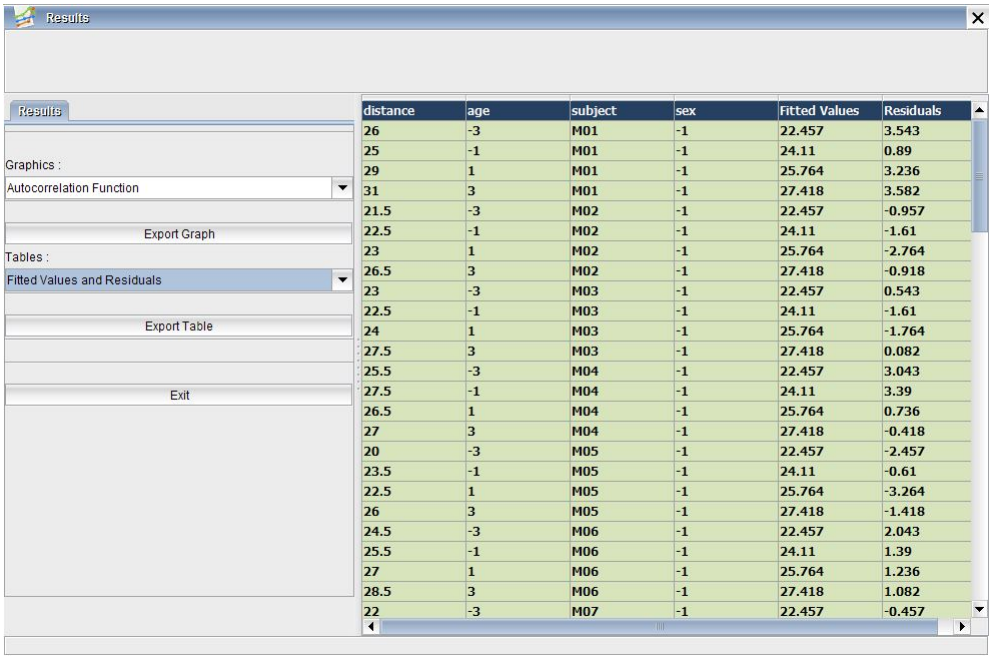


Figure 4.66: Fixed Effects Correlations, GLS



distance	age	subject	sex	Fitted Values	Residuals
26	-3	M01	-1	22.457	3.543
25	-1	M01	-1	24.11	0.89
29	1	M01	-1	25.764	3.236
31	3	M01	-1	27.418	3.582
21.5	-3	M02	-1	22.457	-0.957
22.5	-1	M02	-1	24.11	-1.61
23	1	M02	-1	25.764	-2.764
26.5	3	M02	-1	27.418	-0.918
23	-3	M03	-1	22.457	0.543
22.5	-1	M03	-1	24.11	-1.61
24	1	M03	-1	25.764	-1.764
27.5	3	M03	-1	27.418	0.082
25.5	-3	M04	-1	22.457	3.043
27.5	-1	M04	-1	24.11	3.39
26.5	1	M04	-1	25.764	0.736
27	3	M04	-1	27.418	-0.418
20	-3	M05	-1	22.457	-2.457
23.5	-1	M05	-1	24.11	-0.61
22.5	1	M05	-1	25.764	-3.264
26	3	M05	-1	27.418	-1.418
24.5	-3	M06	-1	22.457	2.043
25.5	-1	M06	-1	24.11	1.39
27	1	M06	-1	25.764	1.236
28.5	3	M06	-1	27.418	1.082
22	-3	M07	-1	22.457	-0.457

Figure 4.67: Adjusted Values and Residuals, GLS

Wald statistic, which is defined as the ratio between the square radius regression and its standard error. This statistic follows a χ^2 with 1 degree of freedom. We can see that **age** has an impact of 0.652 per each time unit; and this change is statistically significant at a level of $\alpha = 0.05$. That is, that for both groups (male and female), the the average growth is 0.652 mm per year, around 11 years. Since the contrast type for **sex** was *Helmert*; one effect were defined as: **sex1** the effect that represents the difference between the means of the observations of the two groups. In our case, there is a significant difference in the distance from the pituitary to pterygomaxilar fissure between men and women. Then the man's contribution to distance is 1.136, while that of a woman is -1.136. Now to see if growths between Men and women are different, we focus on the interaction **age:sex1**. The effect of this interaction is statistically significant at $\alpha = 0.05$ level. This means that the difference between the growth rates of men and women around age of 11 years are different

- **Residuals Normal Plot.** This graph shows residuals empirical quantiles and standard normal distribution quantiles. It helps us validate whether the residuals are Normal distributed. We want points to forma a diagonal line in the graph. See Figure 4.69.
- **Residuals by Factor Normality Plot.** This graphs shows residuals empirical quantiles and normal standard quantiles for both groups, male and female. It helps us to validate whether the model residuals are distributed Normal in both groups. We want points to forma a diagonal line in both graphs. See Figure 4.70.
- **Residuals and Adjusted Values Plot.** This graph helps us to validate the assumption of constant variance in residuals. We note that the dispersion of points in both groups is similar (See Figure 4.71). This is due to the adjustment we made to include a parameter for modeling the difference between the variance of the two groups.
- **Gráfica de Autocorrelación.**

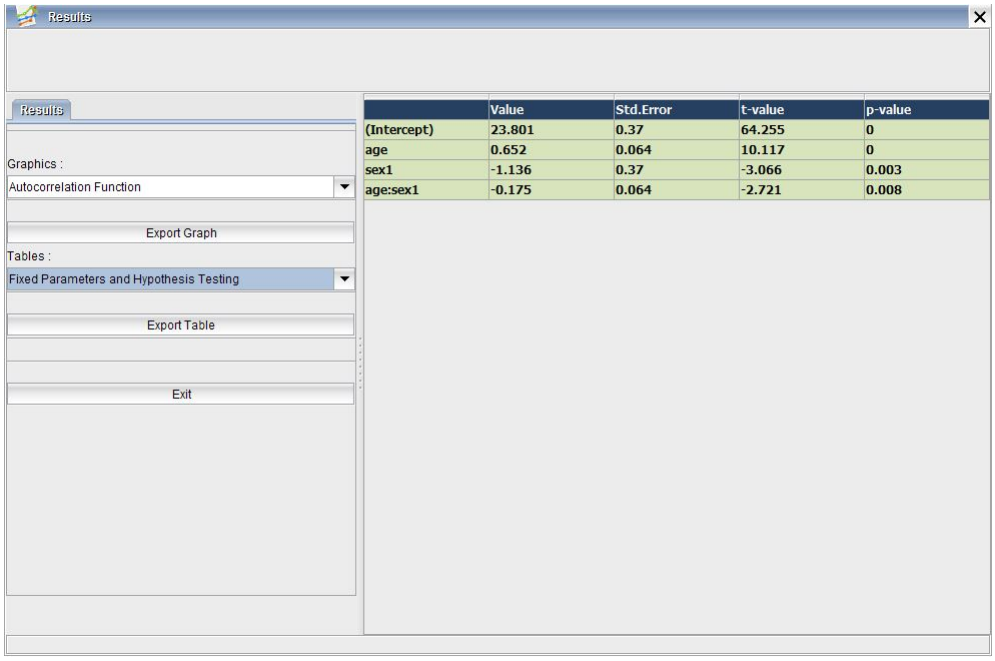


Figure 4.68: Fixed Effects and Hypothesis Testing, GLS

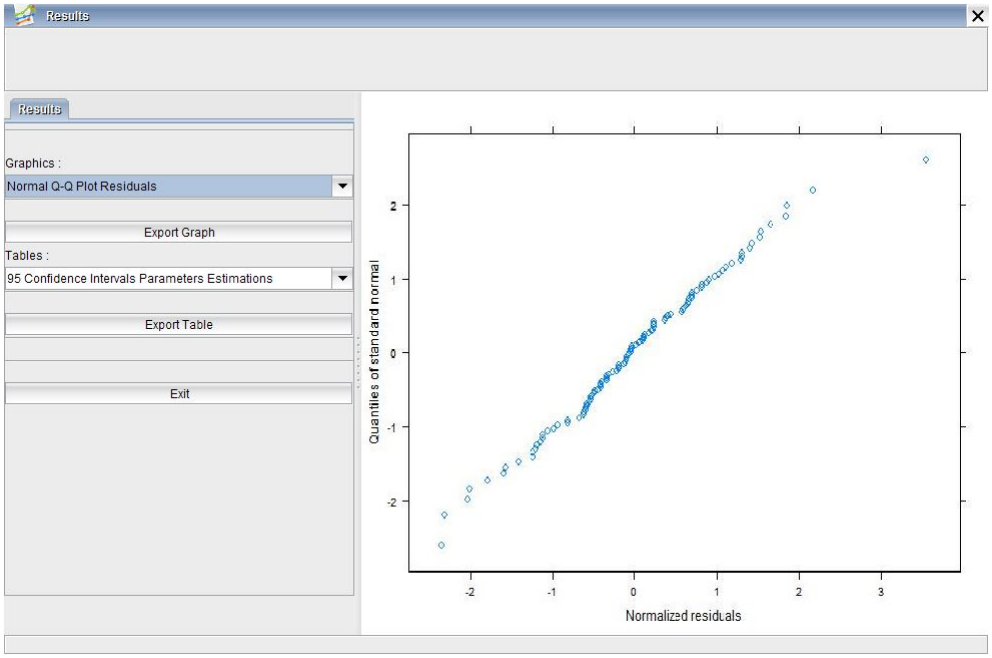


Figure 4.69: Residuals Normal Plot, GLS

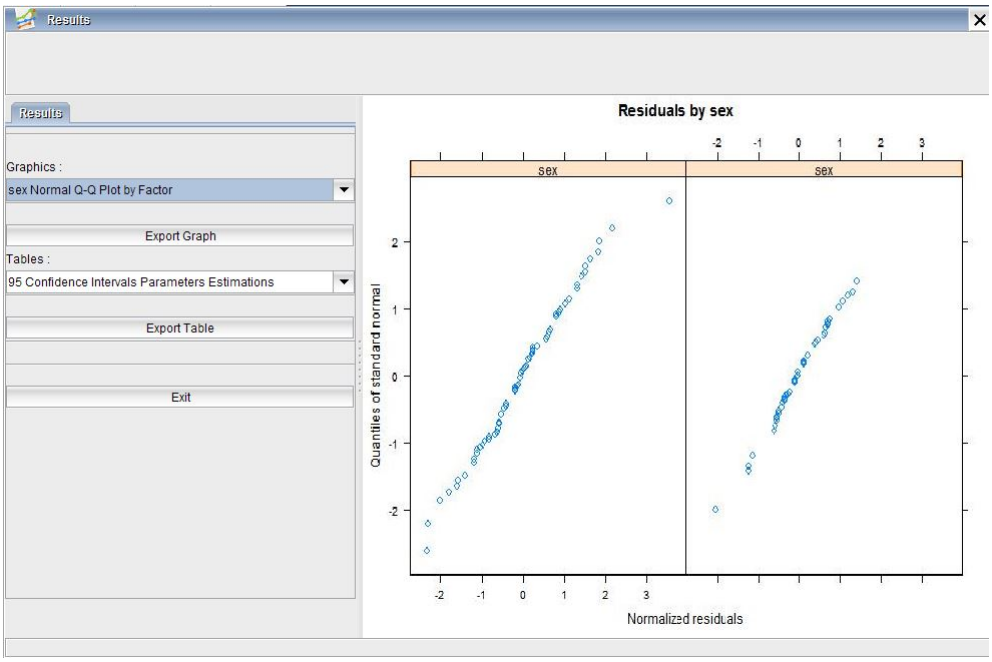


Figure 4.70: Residuals by Factor Normality Plot, GLS

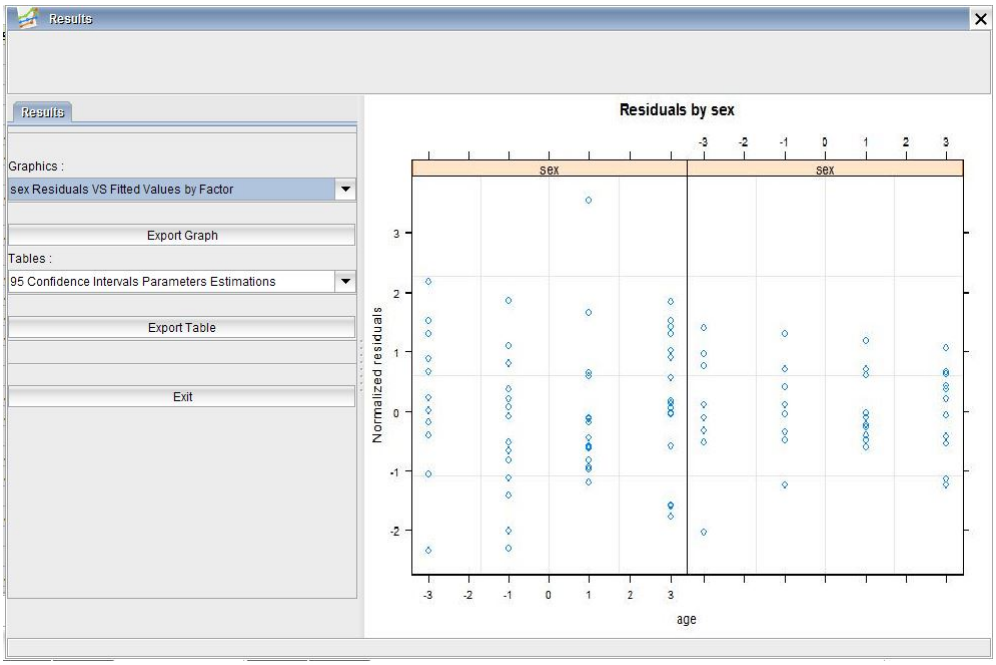


Figure 4.71: Residuals and Adjusted Values Plot, GLS

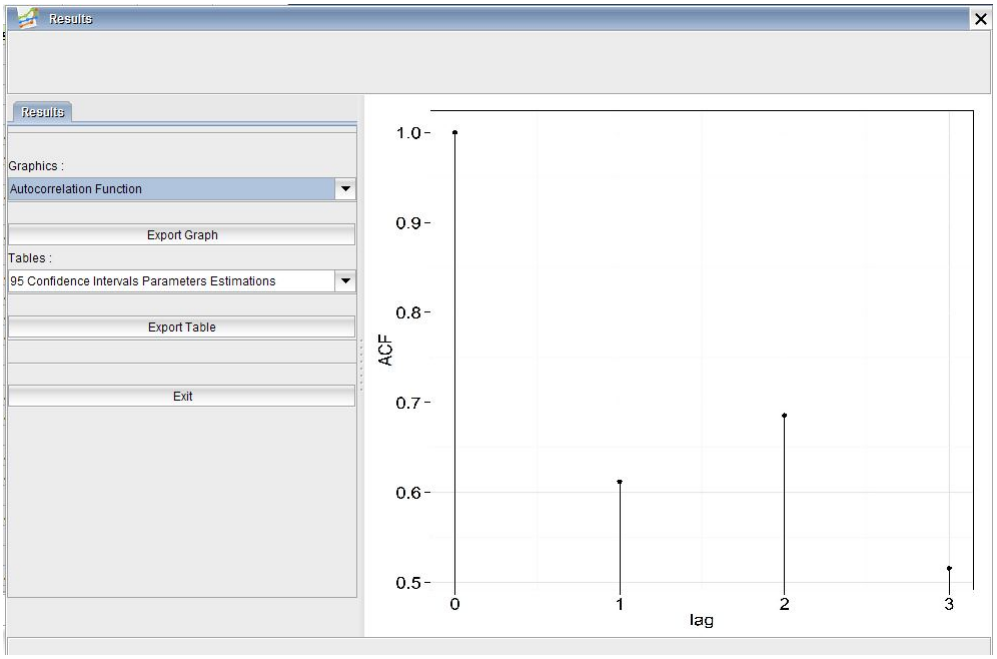
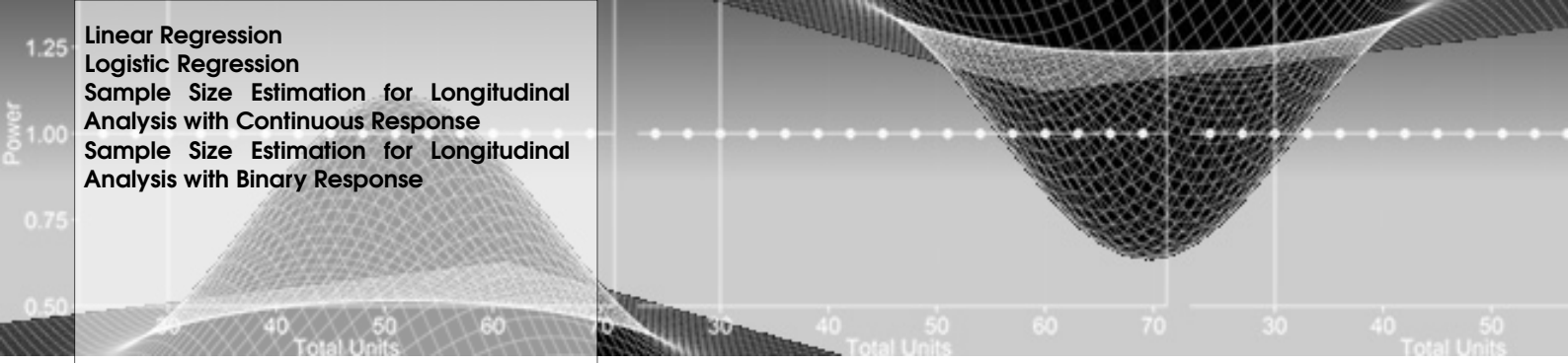


Figure 4.72: Gráfica de residuos en el tiempo por factor, GLS



5 — Statistics Module

5.1 Linear Regression

We will focus on fitting linear regression models to data. To illustrate, suppose you want to develop an empirical model that relates the number of Galapagos tortoises species in several islands to six variables of interest such as: maximum elevation of the Island (elevation, m), area of the island (area, km^2), Distance to the nearest island (Proximity, km^2), The distance to the Santa Cruz island (scruz, km^2), And the area of the adjacent island (adjacent, km^2). A model that can describe this relationship is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon \tag{5.1}$$

where y represents the number of species, x_1 the highest elevation of the island, x_2 the area of the Island, x_3 distance to the nearest island, x_4 the distance to the Santa Cruz island, and x_5 the area to the adjacent island. ε is the vector of errors, and is usually known as residuals. This is a multiple linear regression model with five independent variables. These independent variables are usually known as regressors or predictors. The linear term is due to the form of the parameters β_0, β_1, etc. A nonlinear model would have parameters such as $sin(\beta^2)$, $\log(\beta)$, etc.

There are models that are apparently more complex than (5.1) that can be fitted using linear regression. For example, consider add the interaction between the highest elevation of the island and the its area:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_{12} x_1 x_3 + \varepsilon \tag{5.2}$$

If we set $x_6 = x_1 x_3$, and $\beta_6 = \beta_{13}$; equation (5.1) can be written as follows

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \varepsilon \tag{5.3}$$

and then, we can use linear regression.

Summary of Linear Regression Model

- It is assumed that residuals *i) follow a Normal distribution, 2) have a constant variance, and 3) are independent*. This enables us to derive distributions for model parameters and thus, constructing hypothesis testing to verify its statistical significance. In addition, it allows us to give more precise parameter estimates.
- The method of **least squares** is used to estimate the parameters $\beta_0, \ldots, \beta_k$. The method of least squares finds the values (estimates) of $\beta_0, \ldots, \beta_5$, such that the error (the difference between the responses y and the predictive values of the model, 5.1) is minimized. If the conditions of paragraph 1 are met, then the estimators are unbiased and have minimum variance.
- The **Analysis of Variance** is used as one way to test the hypothesis about the significance of the regressors. Basically, a statistical test is built using the ratio of the contribution made by the regressor x_i through β_i and a unbiased estimate of the variance of the residuals. This test statistic is compared with the F distribution.

Analysis using Linear Regression

To access this function, we simply follow the route **Statistics** $->$ **Linear Regression**. We note that we have previously (see Figure 5.1) loaded the data using **Import**. This data is available in **galapagos.csv**.

We add the variables that conform the model we want to adjust as our first approximation; model (5.1). Figure 5.2 shows the LADES' screen where we have to input the corresponding variables. To adjust proposed model we click on **Set**.

After fitting the model, the results on screen appear; which are shown in Figure 5.3. Next section presents the interpretation and results of this function.

Function Outputs

The tables shown as results are:

- **Hypothesis Testing for Parameters.** Here we use the **t-test** as an equivalent way to test the hypothesis about the significance of the factors included in the model. Of all the factors under study, only **Elevation** and **Adjacent** were statistically significant at a level of $\alpha = 0.05$. Thus, these factors have a significant impact on the response variable (Number of species), and the size of this impact is defined by the estimates of the coefficients. See Figure 5.4.
- **Model Matrix.** In this section we can see the model matrix (see A.1) that was used to adjust the multiple linear regression model. Figure 5.5.
- **Estimated Coefficients.** Presents the estimation for each of the coefficients, $\beta_1, \ldots, \beta_5$. Figure 5.6.
- **Fitted Values and Residuals.** Displays a table containing the fitted values (predict) of the model and the residuals (difference of the actual and predicted). Figure 5.7.
- **Adjusted R-square, Multiple R-square and Residual's Standard Error.** Multiple R-square indicates the proportion of variation in the dependent variable explained by the predictor or independent variables. These two statistics help us to check the adequacy or fit of the proposed model. For our case, it was showed that Multiple R-square = 0.765 and Adjusted R-square = 0.7171. The closer they are to the 1, the better the fit. See Figure 5.8. See Figure 5.8.
- **Analysis of Variance (ANOVA).** Displays the sum of squares for each one of the dependent variables as well as the construction of each of the F-tests (Figure 5.9). The variables

Island	NS	ES	Area	Anear	Dist	DistSC	Elevation
Baltra	58	23	25.09	1.84	0.6	0.6	346
Bartolome	31	21	1.24	572.33	0.6	26.3	109
Caldwell	3	3	0.21	0.78	2.8	58.7	114
Champion	25	9	0.1	0.18	1.9	47.4	46
Coamano	2	1	0.05	903.82	1.9	1.9	119
Daphne_M...	18	11	0.34	1.84	8	8	93
Darwin	10	7	2.33	2.85	34.1	290.2	168
Eden	8	4	0.03	17.95	0.4	0.4	71
Enderby	2	2	0.18	0.1	2.6	50.2	112
Espanola	97	26	58.27	0.57	1.1	88.3	198
Fernandina	93	35	634.49	4669.32	4.3	95.3	1494
Gardner1	58	17	0.57	58.27	1.1	93.1	49
Gardner2	5	4	0.78	0.21	4.6	62.2	227
Genovesa	40	19	17.35	129.49	47.4	92.2	76
Isabela	347	89	4669.32	634.49	0.7	28.1	1707
Marchena	51	23	129.49	59.56	29.1	85.9	343
Onslow	2	2	0.01	0.1	3.3	45.9	25
Pinta	104	37	59.56	129.49	29.1	119.6	777
Pinzon	108	33	17.95	0.03	10.7	10.7	458
Las_Plazas	12	9	0.23	25.09	0.5	0.6	94
Rabida	70	30	4.89	572.33	4.4	24.4	367
San_Cristobal	280	65	551.62	0.57	45.2	66.5	716
San_Salvador	237	81	572.33	4.89	0.2	19.8	906
Santa_Cruz	444	95	903.82	0.52	0.6	0	864
Santa_Fe	62	28	24.08	0.52	16.5	16.5	259
Santa_Maria	285	73	170.92	0.1	2.6	49.2	640
Seymour	44	16	1.84	25.09	0.6	9.6	147

Figure 5.1: Loaded Data

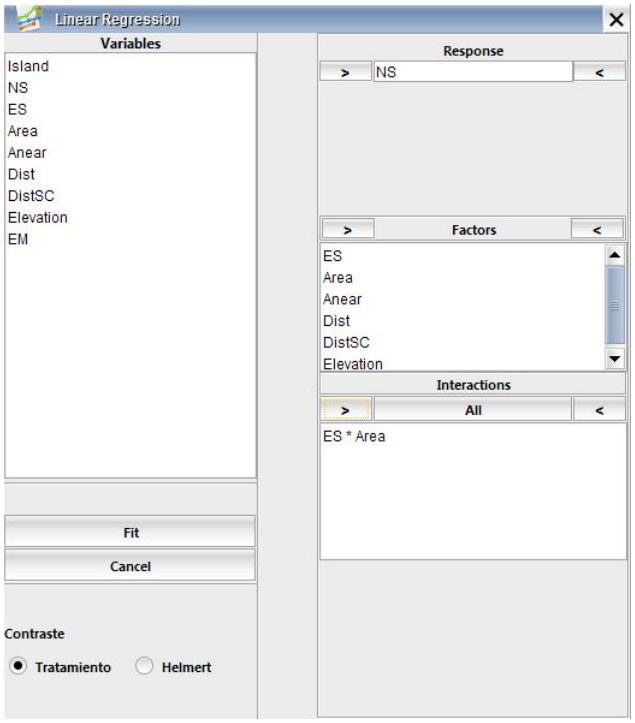


Figure 5.2: Linear Regression Main Screen

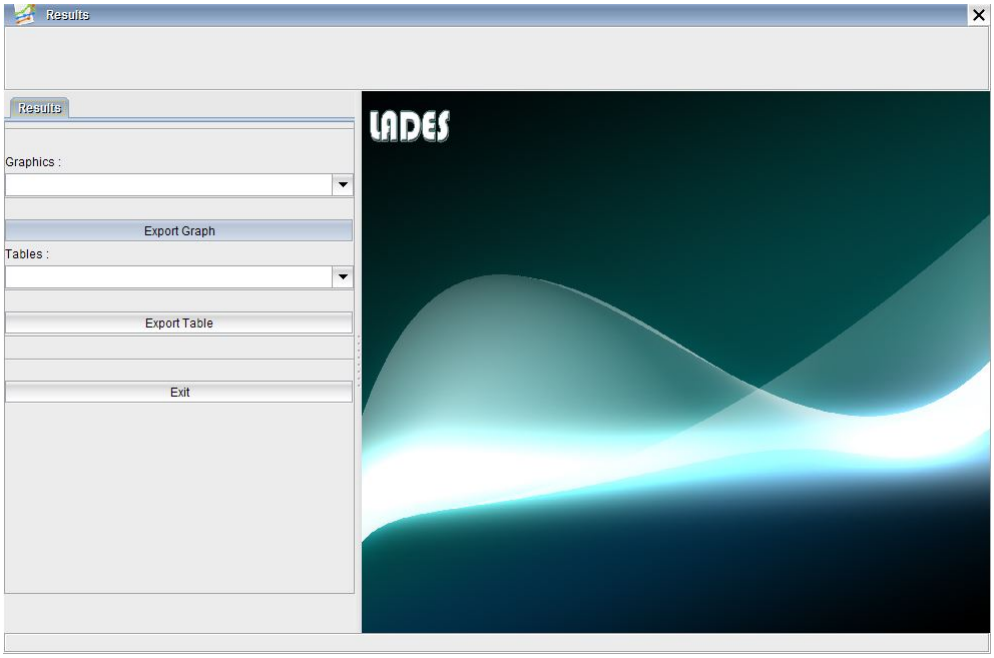
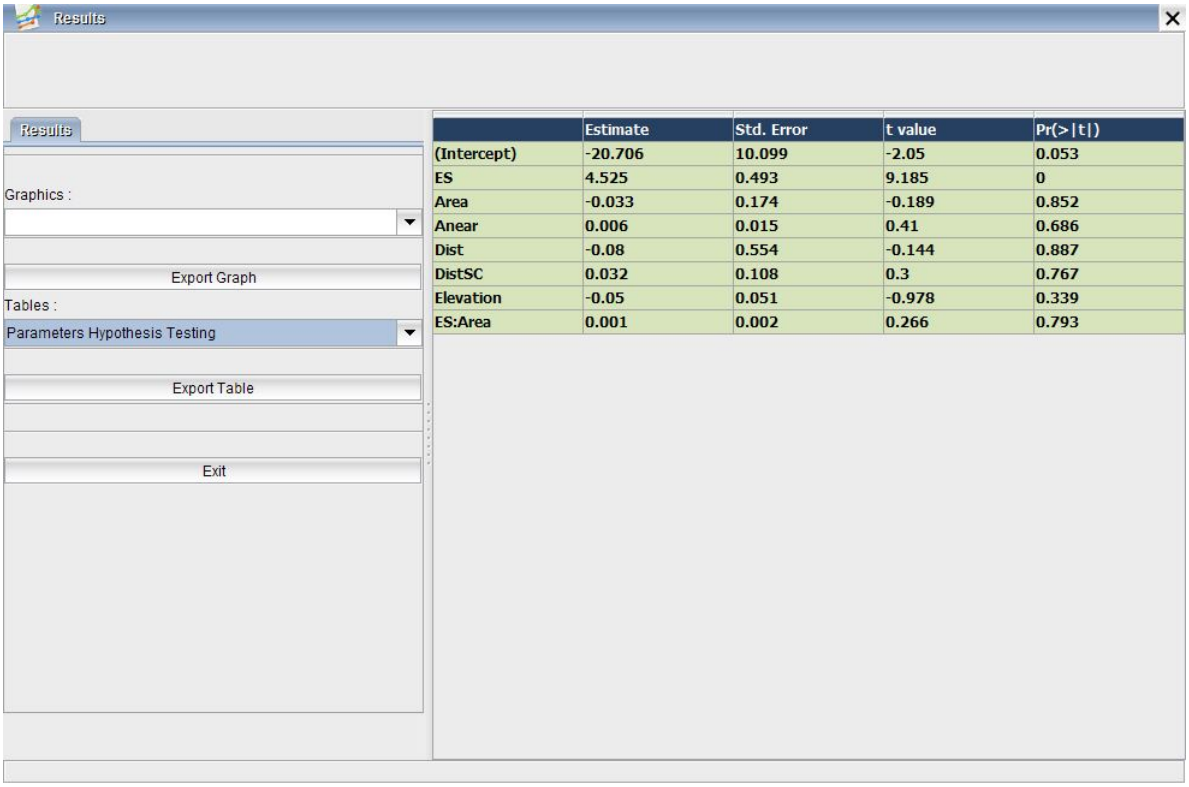
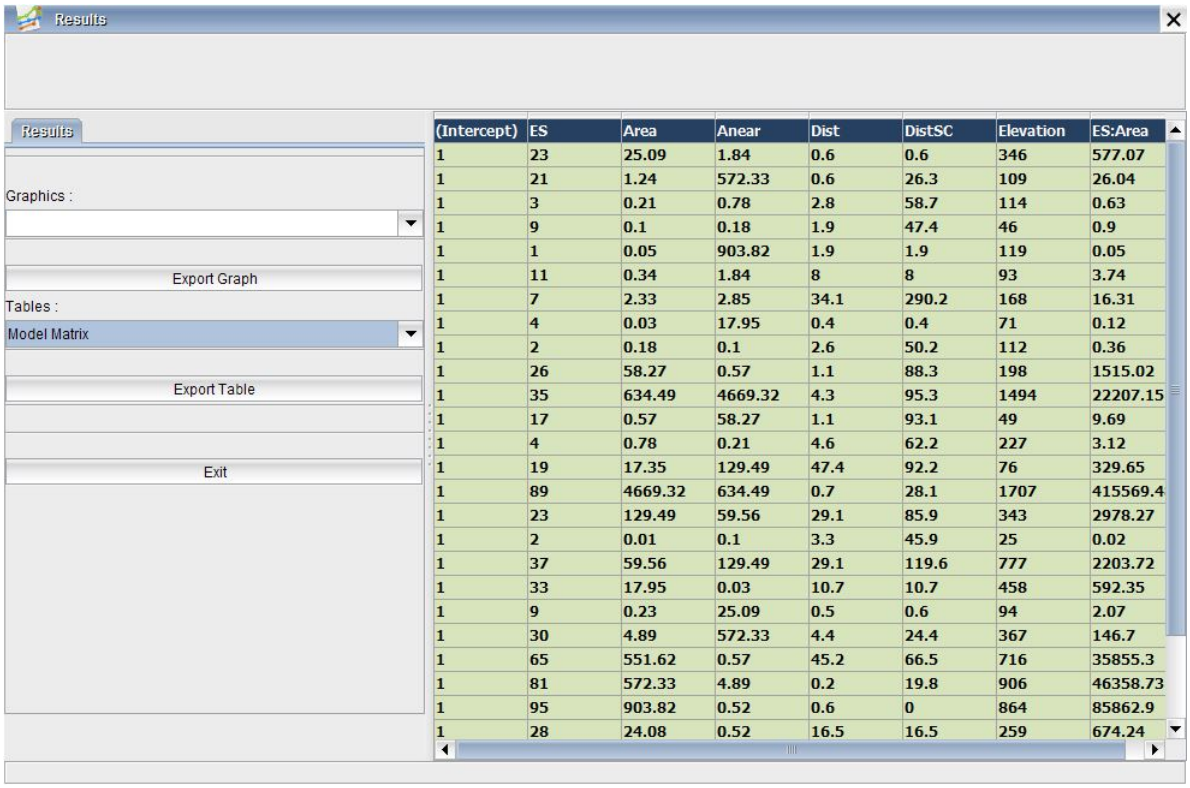


Figure 5.3: Results Screen, Linear Regression



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-20.706	10.099	-2.05	0.053
ES	4.525	0.493	9.185	0
Area	-0.033	0.174	-0.189	0.852
Anear	0.006	0.015	0.41	0.686
Dist	-0.08	0.554	-0.144	0.887
DistSC	0.032	0.108	0.3	0.767
Elevation	-0.05	0.051	-0.978	0.339
ES:Area	0.001	0.002	0.266	0.793

Figure 5.4: Hypothesis Testing for Parameters, Linear Regression



(Intercept)	ES	Area	Anear	Dist	DistSC	Elevation	ES:Area
1	23	25.09	1.84	0.6	0.6	346	577.07
1	21	1.24	572.33	0.6	26.3	109	26.04
1	3	0.21	0.78	2.8	58.7	114	0.63
1	9	0.1	0.18	1.9	47.4	46	0.9
1	1	0.05	903.82	1.9	1.9	119	0.05
1	11	0.34	1.84	8	8	93	3.74
1	7	2.33	2.85	34.1	290.2	168	16.31
1	4	0.03	17.95	0.4	0.4	71	0.12
1	2	0.18	0.1	2.6	50.2	112	0.36
1	26	58.27	0.57	1.1	88.3	198	1515.02
1	35	634.49	4669.32	4.3	95.3	1494	22207.15
1	17	0.57	58.27	1.1	93.1	49	9.69
1	4	0.78	0.21	4.6	62.2	227	3.12
1	19	17.35	129.49	47.4	92.2	76	329.65
1	89	4669.32	634.49	0.7	28.1	1707	415569.4
1	23	129.49	59.56	29.1	85.9	343	2978.27
1	2	0.01	0.1	3.3	45.9	25	0.02
1	37	59.56	129.49	29.1	119.6	777	2203.72
1	33	17.95	0.03	10.7	10.7	458	592.35
1	9	0.23	25.09	0.5	0.6	94	2.07
1	30	4.89	572.33	4.4	24.4	367	146.7
1	65	551.62	0.57	45.2	66.5	716	35855.3
1	81	572.33	4.89	0.2	19.8	906	46358.73
1	95	903.82	0.52	0.6	0	864	85862.9
1	28	24.08	0.52	16.5	16.5	259	674.24

Figure 5.5: Model Matrix, Linear Regression

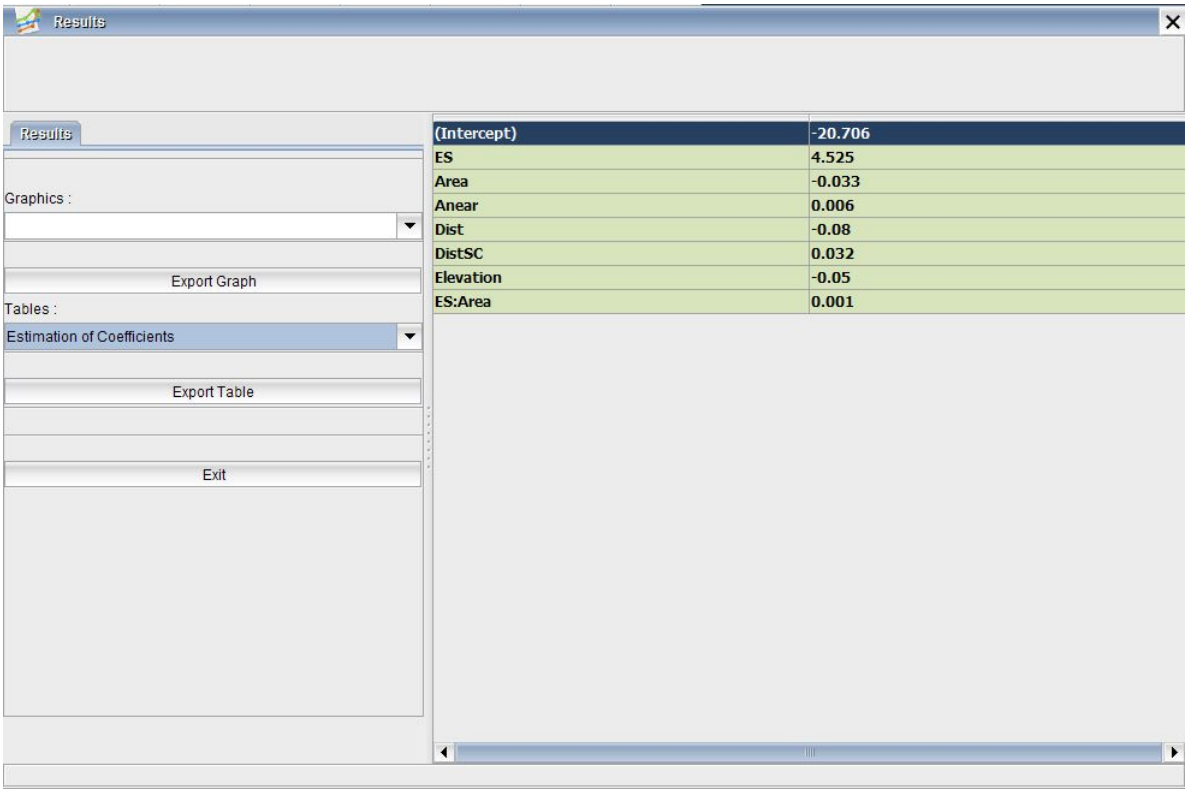


Figure 5.6: Estimated Coefficients

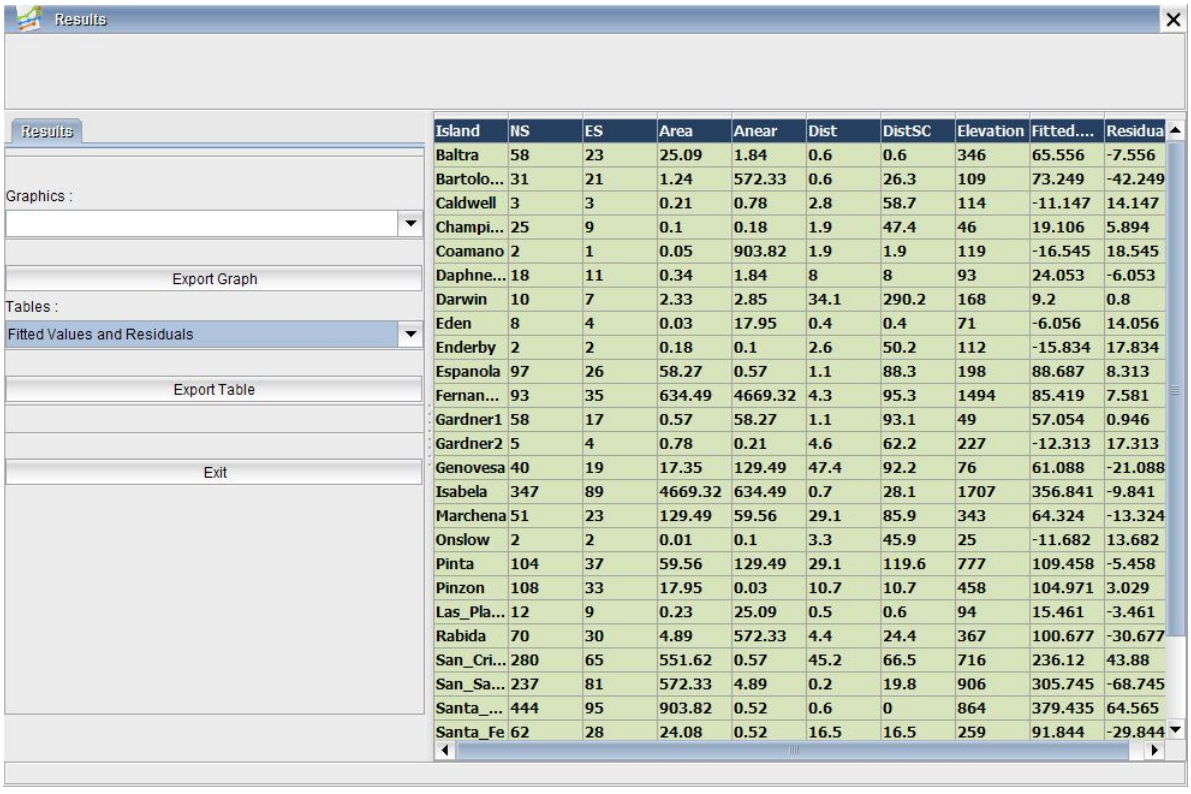


Figure 5.7: Fitted Values and Residuals

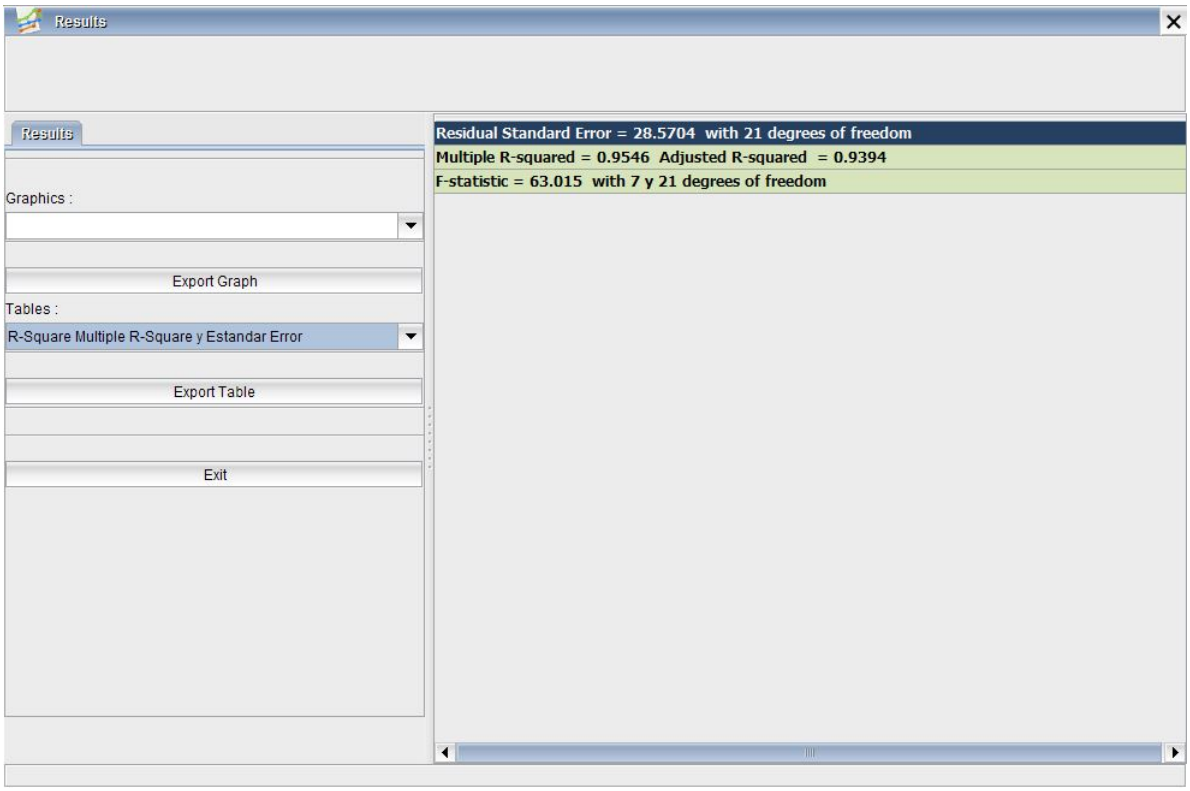
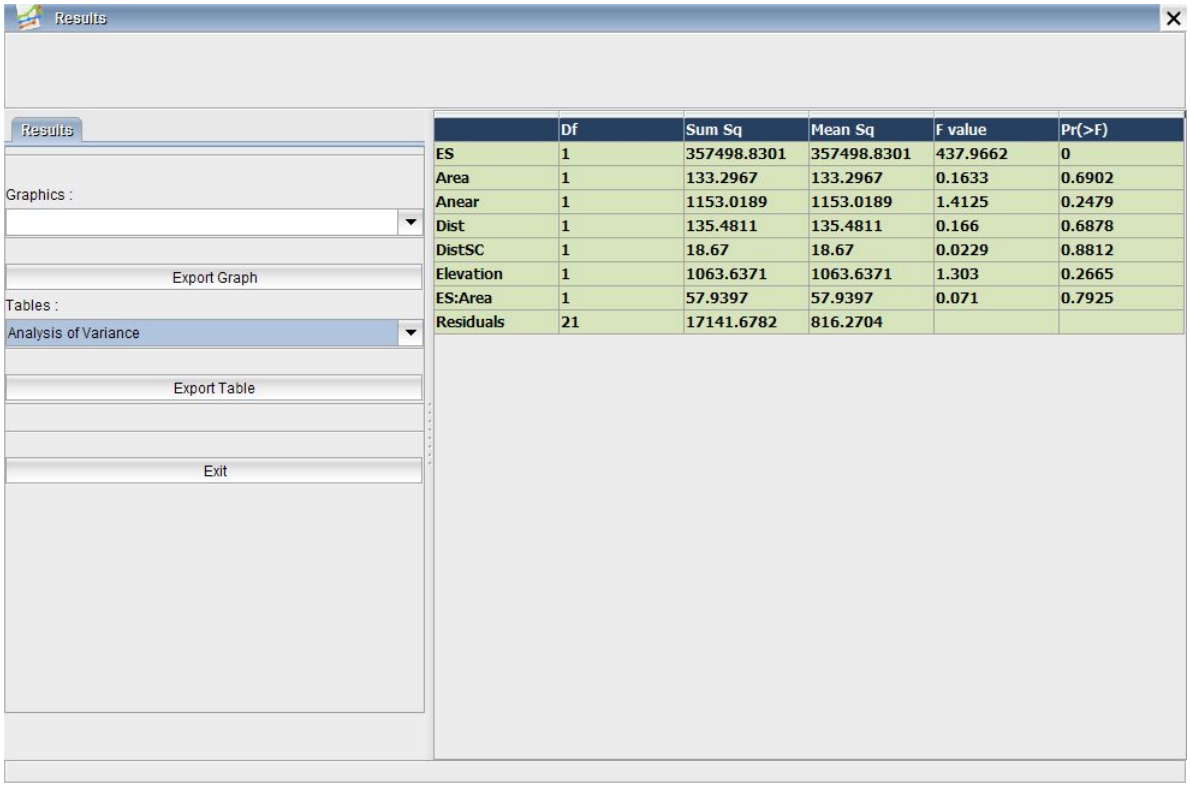

Results
Results
Graphics :
Export Graph
Tables :
R-Square Multiple R-Square y Estandar Error
Export Table
Exit
Residual Standard Error = 28.5704 with 21 degrees of freedom
Multiple R-squared = 0.9546 Adjusted R-squared = 0.9394
F-statistic = 63.015 with 7 y 21 degrees of freedom

Figure 5.8: Adjusted R-square, Multiple R-square and Residual's Standard Error, Linear Regression.

Area, **Elevation** and **Adjacent** are significant to a level of $\alpha = 0.05$. The variable **Area** was not significant in our first hypothesis testing; using the t-test. However, ANOVA considered as the most reliable test to determine the significance factors. Therefore, we have three significant factors that affect the number of species.



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ES	1	357498.8301	357498.8301	437.9662	0
Area	1	133.2967	133.2967	0.1633	0.6902
Anear	1	1153.0189	1153.0189	1.4125	0.2479
Dist	1	135.4811	135.4811	0.166	0.6878
DistSC	1	18.67	18.67	0.0229	0.8812
Elevation	1	1063.6371	1063.6371	1.303	0.2665
ES:Area	1	57.9397	57.9397	0.071	0.7925
Residuals	21	17141.6782	816.2704		

Figure 5.9: ANOVA, Linear Regression

- **Residuals Histogram.** Shows the histogram of the residuals. It helps us validate the assumption of *normality, i)* (Figure 5.10). The graph should, preferably, reflect a standard normal distribution, a bell shape centered at zero. For our data, this assumption seems to be fulfilled.
- **Normal QQ Plot.** This plot shows the empirical quantiles for residuals and the standard normal distribution quantiles. It helps to validate the assumption of *normality, i)*. We want the points on the graph to be located close to the midline. This chart helps us to detect possible 'outliers' that occurred in our study. For our data, we can see that most falls around the midline. We can detect, also an atypical point in the top right of the graph. See Figure 5.11.
- **Independence of Residuals Plot.** This graph shows the Times (X axis) in which each observation was collected, and their corresponding residuals (Y axis). It helps us to validate the assumption of *independence, iii)*. We want the plotted line to show no trend up or down, ii.e. to be around around zero. If you have an upward trend, for example, this would give indication that the result of the current observation, depends on the latter. For our data, it appears that there is no explicit trend. See Figure 5.12.
- **Residuals VS Fitted Values Plot.** Fitted values are plotted on the X axis, and their corresponding residuals in the Y axis. It helps us to validate the *constant variance assumption ii)*. It is desired that points on the graph to form an horizontal band. If the points on the graph reflect a form of cone; then, the variance of the residuals is not constant.

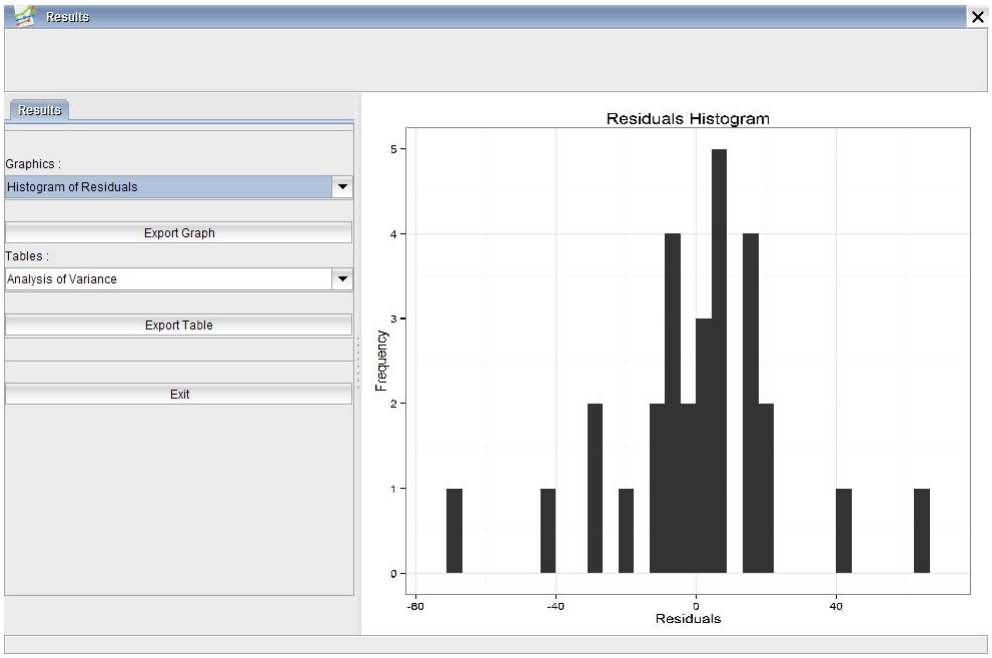


Figure 5.10: Residuals Histogram, Linear Regression

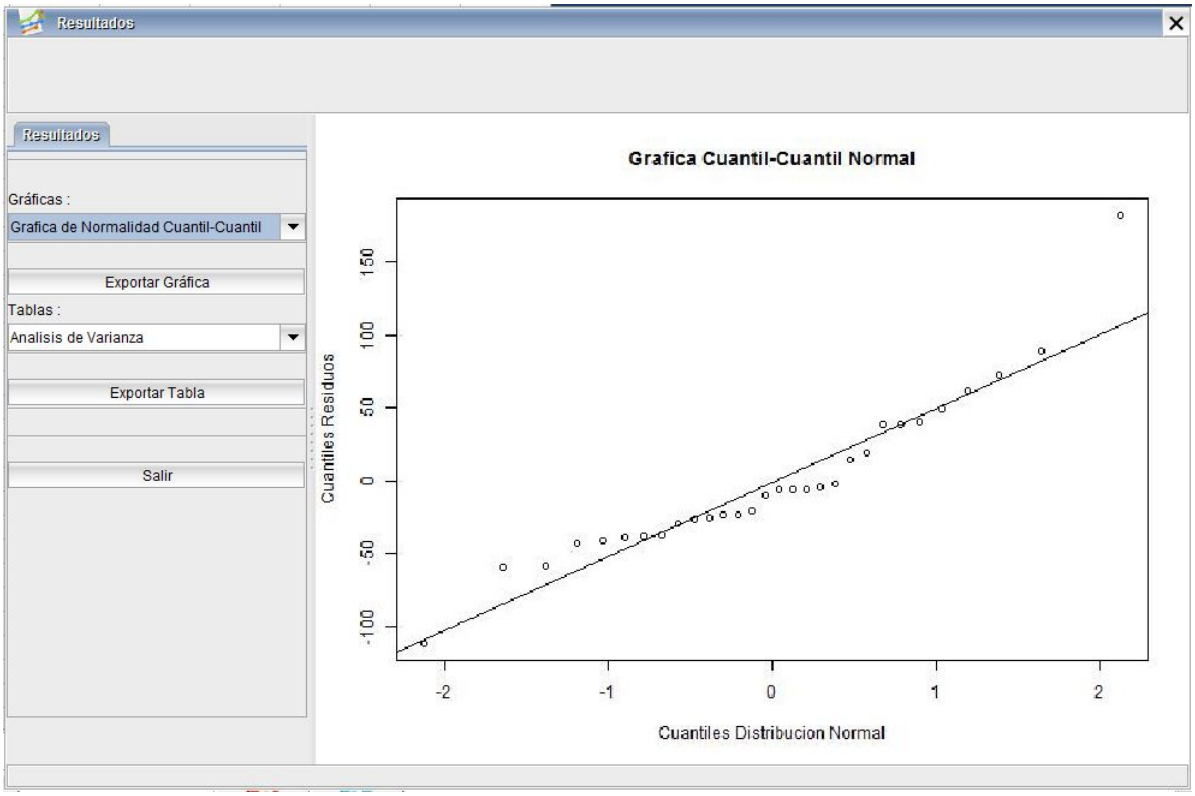


Figure 5.11: Normal QQ Plot, Linear Regresion

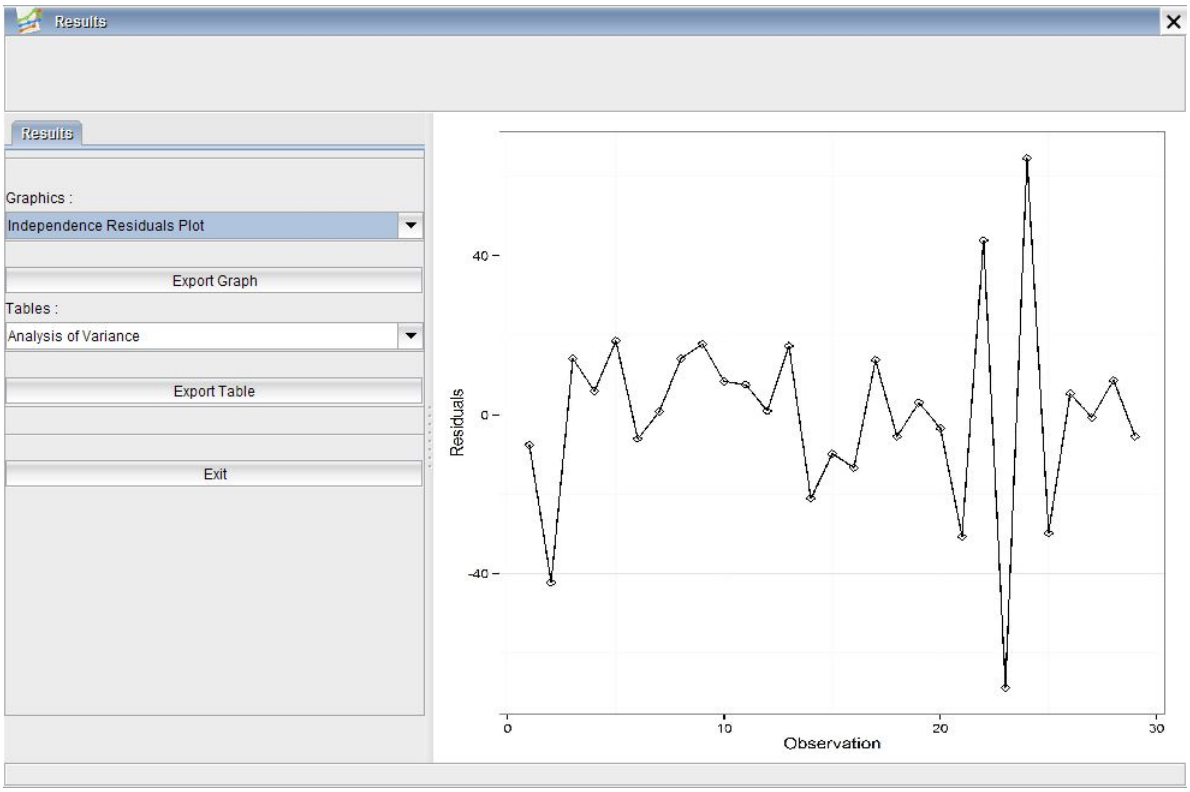


Figure 5.12: Independence of Residuals Plot

For our data, we can see a form of "cone" in the graph (Figure 5.13) which indicates that assumption ii) is not satisfied and we must set another model with other features or predictors.

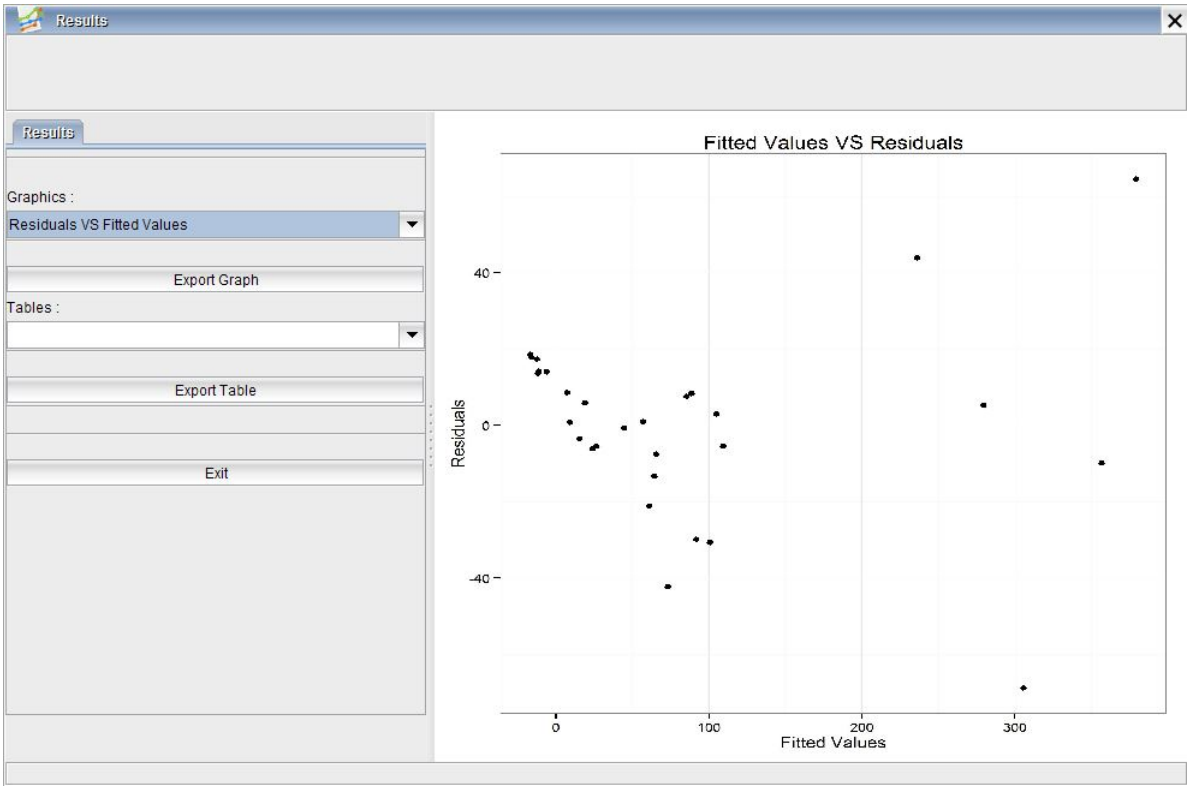


Figure 5.13: Residuals VS Fitted Values Plot

5.2 Logistic Regression

The logistic regression develops the special case in which the variable of interest only takes two values, e.g. present or not present. Since the response can have two values, we can always encode the two cases as 0 and 1. For example no present = 0, present = 1;ll for certain types of disease in an individual. Then, the probability p that the response is equal to 1, i.e. the probability that a patient has the disease, is of interest. The distribution for binary type variables is the **Bernoulli** distribution. This distribution has the following properties

- Mean = p
- Variance = $p(1-p)$

We can clearly see that the variance depends upon p.

Since the response of interest can only have values two, 0 or 1; if we want to model the probability that y is equal to 1 for a single predictor, we can write the following model

$$p = E(Y|z) = \beta_0 + \beta_1 z$$

and add the error of the model ε. Two problems arise with such type of modeling:

1. The predicted values by our model for response Y may be greater than 1 or less than 0 since the linear expression for these predictions is not restricted
2. One of the assumptions of linear regression analysis is that variance of Y is constant throughout the values of the predictor z. We have shown that this is not the case.

Summary of logistic regression model

- A new response is constructed to transform the current response as follows

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = \eta \tag{5.4}$$

where η is the predicted value by the model, e.g. $\eta = \beta_0 + \beta_1 = z$. This function of the probability p is called **logit**. This function no longer has restrictions and we can work freely on it, and fit a linear model to it. This model can be extended to more predictors.

- To return to the actual values of variable Y, i.e. restricted values, we use the inverse transformation of (5.2)

$$p = \frac{1}{1+e^{-\eta}}$$

- The estimation of the model parameters (the values of $\beta's$) is done by maximizing the likelihood function. Because there is no closed solution to such estimates, numerical methods are used instead.
- The model parameters, β_k, in his version of inverse transformation, $exp(\beta_i)$, indicate how the probability is changing per unit of z.
- To determine if the effects of our factors are significant to the model, we use the **likelihood ratio test**. The statistic used to this test is as follows

$$-2\ln\left(\frac{L_{max,Reduced}}{L_{max}}\right)$$

That is, we construct an hypothesis $H_0 : \beta_k = 0$. The test statistic is based on the ratio of the $L_{max,Reduced}$ which is the maximum likelihood without including the factor, and L_{max} the maximum likelihood including all the factors. This statistic is also known as *deviance*, and is chi-squared distributed with one degree of freedom. We reject H_0 when the deviance is too large.

- Another hypothesis testing is *Wald* test. The Wald test for $H_0 : \beta_k = 0$ uses the $Z = \hat{\beta}/SE(\hat{\beta})$ statistic, or its chi-squared version Z^2 with one degree of freedom. Where SE represents the Standard Error of the estimation. One requisite for this test is to have a large sample size.

Analysis using Logistic Regression Model

As an example, we have a study in which three types of treatment were analyzed. The response under evaluation was lymphocytic infiltration in mice. In addition to the two treatments under study, a control group was included and a no-Treatment group. The difference between these two is that if the **No-Treatment** group was underwent surgery but without applying any type of treatment. Measurements were taken at 15, 20, 45 and 60 days after infiltrating the disease and apply the treatment. The answers are stored as Present (presence of lymphocytic infiltrate) and Absent.

The data from this experiment are found in the file **infiltration.csv**.

To access the logistic regression function follow the route **Statistics -> Logistic regression**. In Figure 4.14 the main screen is shown function.

In Figure 5.14 we have included the model to be analyzed. The model can be mathematically represented as follows:

$$logit(p) = \beta_0 + \beta_1\, Time + \beta_2\, Treatment + \beta_3\, Time : Treatment + \varepsilon \tag{5.5}$$

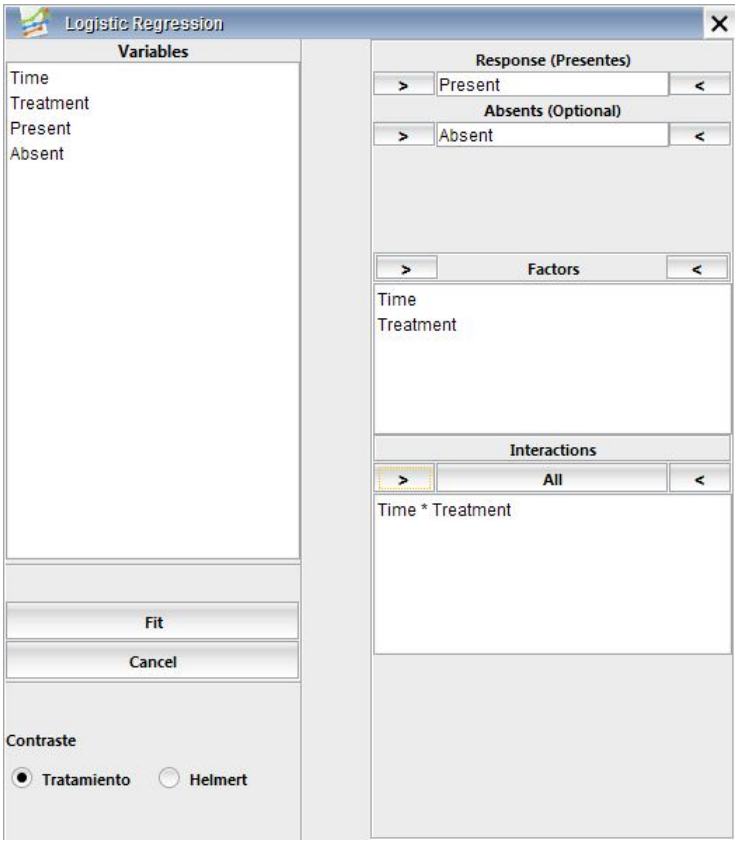
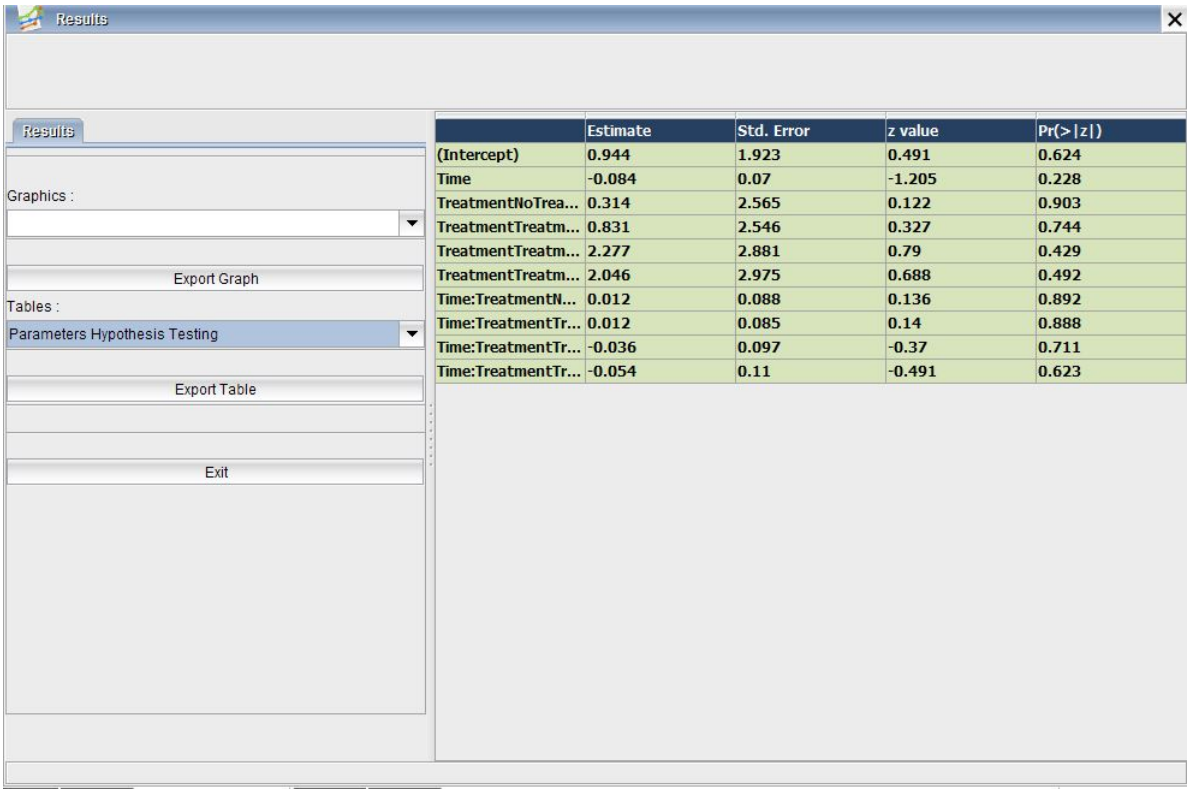


Figure 5.14: Main Screen, Logistic Regression

where p is the proportion of Presents. We want to analyze the effects of Treatments, the effect of time and a possible interaction between these two factors. All samples are independent, so we can work with time as a variable not correlated in its levels. The type of contrast used is *Treatment*, since we have discrete levels for this factor. This means that all results are comparisons against the **Control** group.

Function Outputs

- **Hypothesis Testing for Parameters.** Figure 5.15 shows Wald test for each of the parameters. **Time** has a negative effect of 0.084 in the logit of the response. This effect appears to be significant at a level of $\alpha = 0.05$. For **No infection**, **C++**, **Uu+** and **C+Uu+** treatments, the corresponding coefficients are read as the change that occurs when applying any of these treatments to the control group. For instance, for **C ++** Treatment, a mouse in the control group taking this treatment produces a change in the logit Response of 0.831. We note that no treatment was statistically significant at $\alpha = 0.05$. For the interaction between Treatment and Time, the estimated parameter is: the change in the growth over time of being in the control group and be administrated with one treatment. For example, for the interaction **Time: Treatment Uu +**, this change is 0.012. Neither of these interactions is significant



	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.944	1.923	0.491	0.624
Time	-0.084	0.07	-1.205	0.228
TreatmentNoTrea...	0.314	2.565	0.122	0.903
TreatmentTreatm...	0.831	2.546	0.327	0.744
TreatmentTreatm...	2.277	2.881	0.79	0.429
TreatmentTreatm...	2.046	2.975	0.688	0.492
Time:TreatmentN...	0.012	0.088	0.136	0.892
Time:TreatmentTr...	0.012	0.085	0.14	0.888
Time:TreatmentTr...	-0.036	0.097	-0.37	0.711
Time:TreatmentTr...	-0.054	0.11	-0.491	0.623

Figure 5.15: Hypothesis Testing for Parameters, Logistic Regression

- **Model Matrix.** In this section we can see the model matrix (see A.1) that was used to adjust the logistic regression model. See Figure 5.16.
- **Estimated Coefficients.** Shows the estimation of coefficients $\beta_1, \ldots, \beta_5$. See Figure 5.17. These coefficients are the impacts factors have on the *logit* of the response.
- **Fitted Values and Residuals.** In Figure 5.18, adjusted values and their corresponding residuals are shown. The adjusted values are the resulting values after applying the inverse

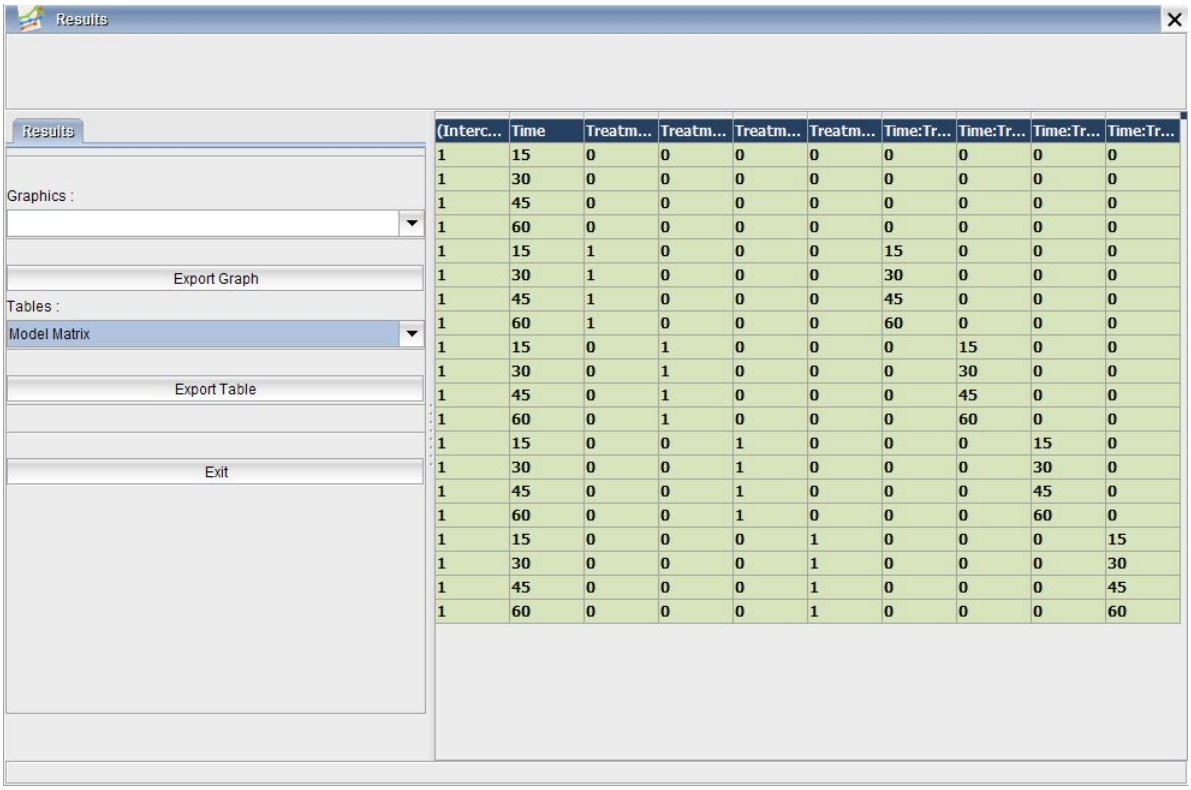


Figure 5.16: Model Matrix, Logistic Regression

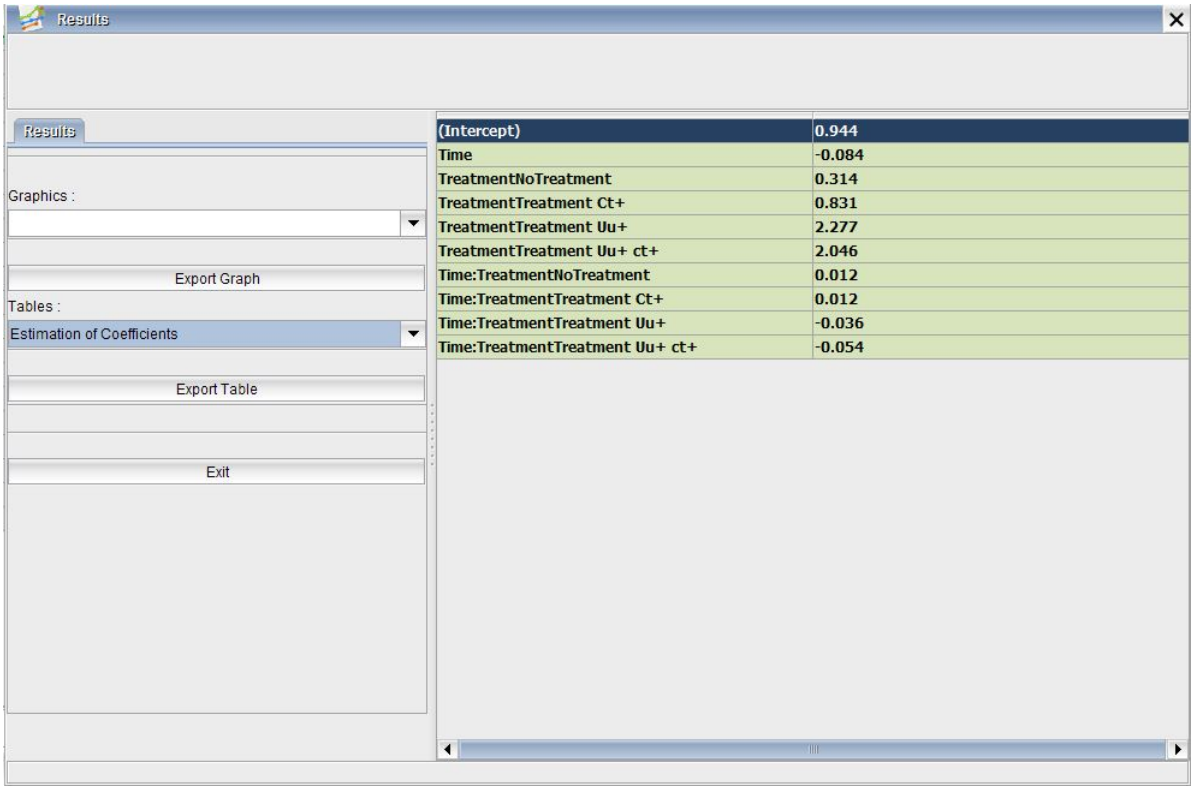
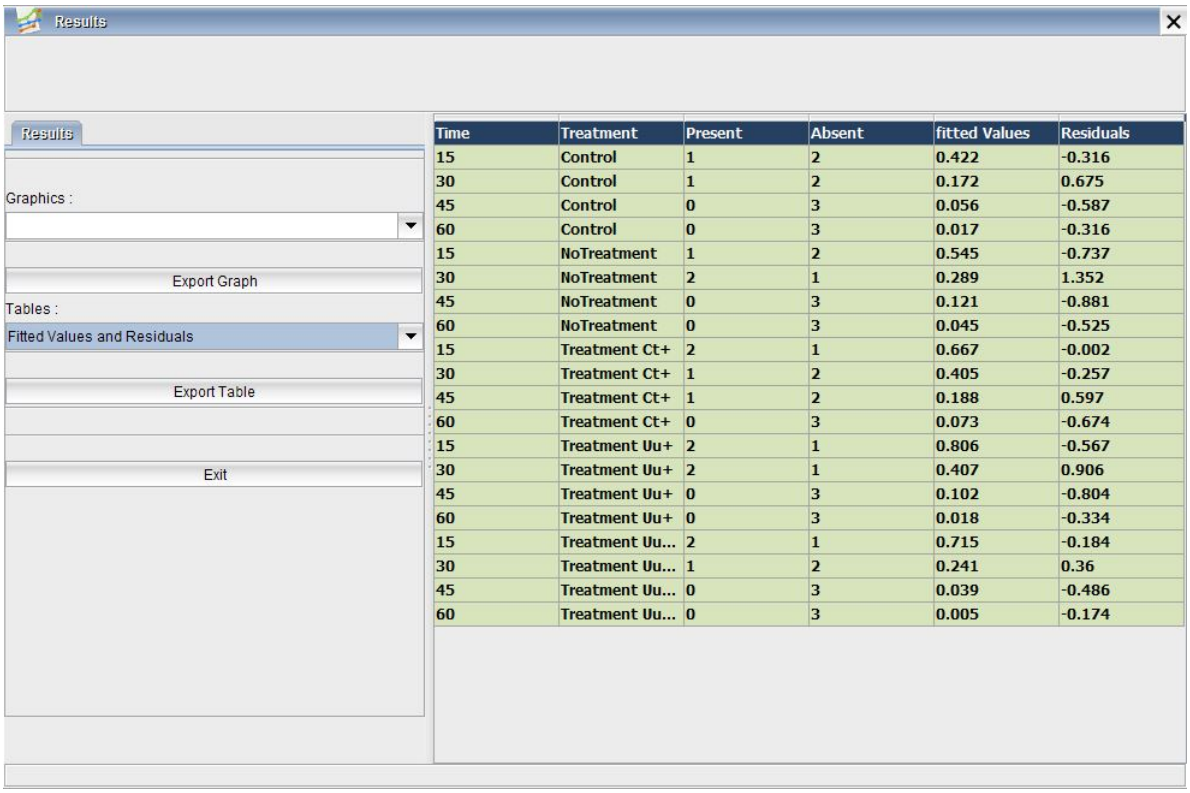


Figure 5.17: Estimated Coefficients, Logistic Regression

transformation. For our case, these values denote the probability that lymphocytic infiltrate is present. The model residuals are unrestricted, i.e. are based on the logit function.



Time	Treatment	Present	Absent	fitted Values	Residuals
15	Control	1	2	0.422	-0.316
30	Control	1	2	0.172	0.675
45	Control	0	3	0.056	-0.587
60	Control	0	3	0.017	-0.316
15	NoTreatment	1	2	0.545	-0.737
30	NoTreatment	2	1	0.289	1.352
45	NoTreatment	0	3	0.121	-0.881
60	NoTreatment	0	3	0.045	-0.525
15	Treatment Ct+	2	1	0.667	-0.002
30	Treatment Ct+	1	2	0.405	-0.257
45	Treatment Ct+	1	2	0.188	0.597
60	Treatment Ct+	0	3	0.073	-0.674
15	Treatment Uu+	2	1	0.806	-0.567
30	Treatment Uu+	2	1	0.407	0.906
45	Treatment Uu+	0	3	0.102	-0.804
60	Treatment Uu+	0	3	0.018	-0.334
15	Treatment Uu...	2	1	0.715	-0.184
30	Treatment Uu...	1	2	0.241	0.36
45	Treatment Uu...	0	3	0.039	-0.486
60	Treatment Uu...	0	3	0.005	-0.174

Figure 5.18: Fitted Values and Residuals, Logistic Regression

- **Analysis of Variance, ANOVA.** This analysis of variance works different as ANOVA for linear regression. This typo of ANOVA use the likelihood ration for testing hypothesis over factors. For our example, only **Time** was significant; that is, adding Time to our proposed model contributes significantly to the explanation of the observed data. This hypothesis testing works with small samples.
- **Null Deviance and Residual.** Null deviance of the fitted model. See Figure 5.20.

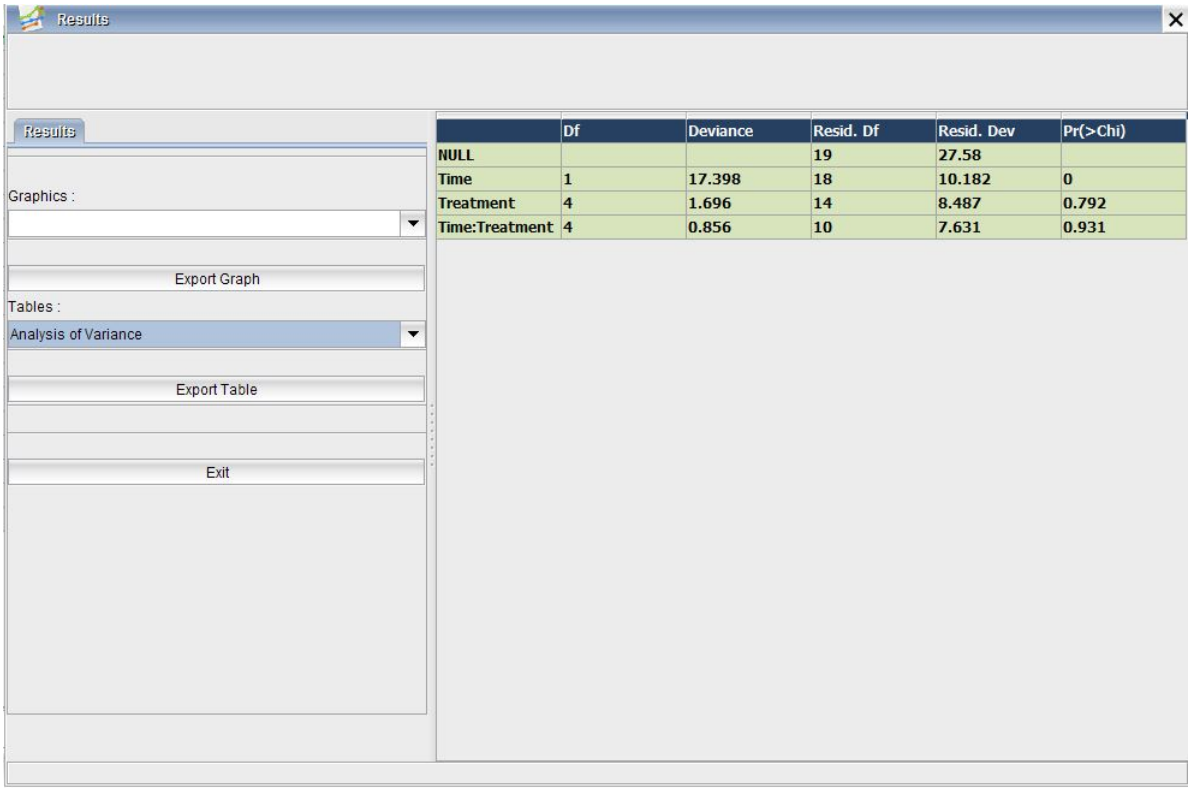


Figure 5.19: ANOVA, Logistic Regression

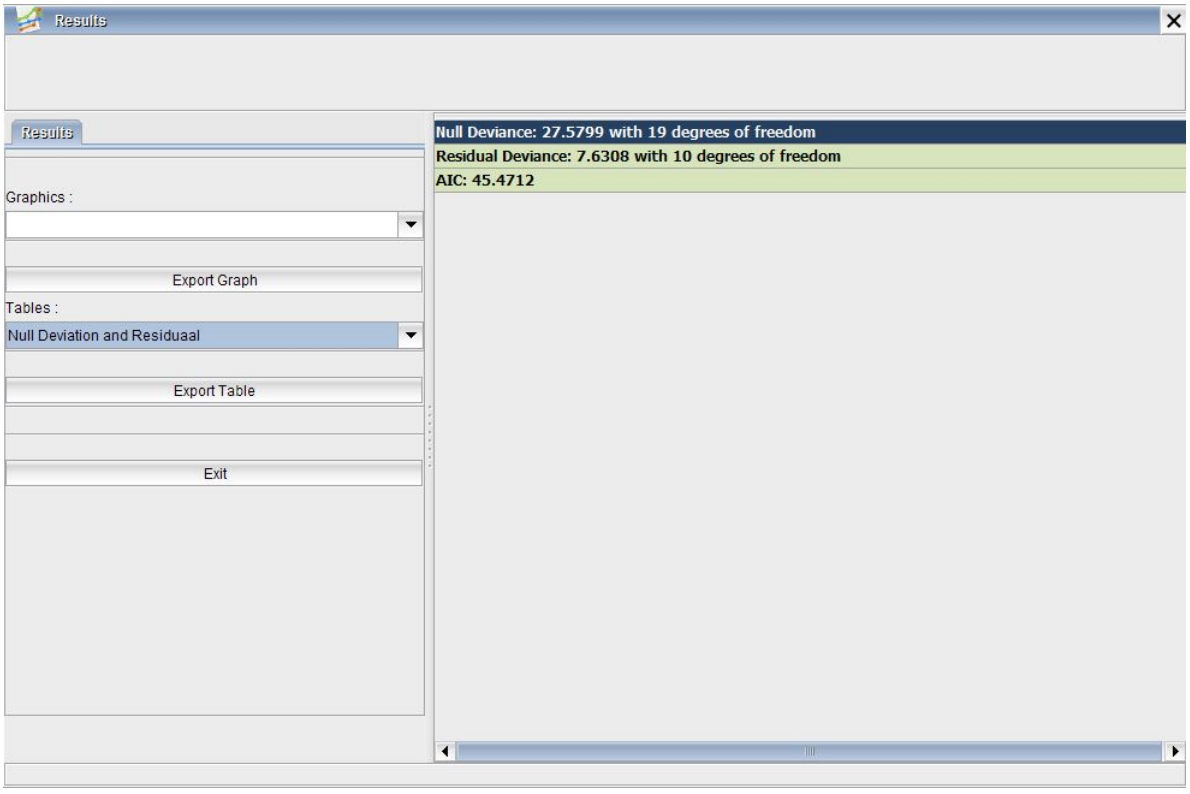


Figure 5.20: Null Deviance and Residual, Logistic Regression

5.3 Sample Size Estimation for Longitudinal Analysis with Continuous Response

Before conducting a longitudinal study is necessary to calculate the number of subjects needed to ensure that certain predefined effect is significant. Sample size calculation are based on assumptions which, if they are not satisfied, can lead to very different sizes. Sample size calculation is related to level of significance (How many individuals are needed to make some effect significant?).

As an example, we have a study in which we want to compare a therapy-based intervention against a placebo intervention. In this study the response of interest is the systolic pressure. Previous data collected for the sample size calculation of this study are shown in 5.1.

Information	**Value**
Average Standard Deviation in the first and second measurement	13.6
Time Measures Correlation	0.867
Number of measures over Time	2
Significance Level	0.05
Power	0.8
Significant Difference	4.23

Table 5.1: Previous data needed to compute sample size for a continuos response

Suppose we have the following model for the Control group

$$Y_{ij} = \beta_{0C} + \beta_{1C} t_{ij} + \varepsilon_{ij} \tag{5.6}$$

where Y_{ij} are the observations of patient i in time j. β_{0C} represents the intercept of the control group; and β_{1C} represents the average changing rate over time. The same model can be constructed for the Treatment group. The sample size is based on detecting a significant difference between the average change of the two groups in time, i.e. β_{1C} y β_{1T}.

Summary of Sample Size Estimation

- The null hypothesis on which sample size calculations are based is: $H_0 : \beta_{1C} = \beta_{1T}$. That is, the average change over time for the two groups is the same. The alternative hypothesis can be of two types $H_0 : \beta_{1C} \neq \beta_{1T}$ (two-sided) or $H_0 : \beta_{1C} > \beta_{1T}$ (one sided) (or, $H_0 : \beta_{1C} < \beta_{1T}$).
- The required prior information to make the calculations is:

  α Type 1 Error. This parameter represents the probability the null hypothesis is rejected when is true, for example, it would correspond to the probability of concluding there is a difference between treatment and control group when there is not. Usually equal to 0.05.

  δ Minimum difference to detect. Researchers usually want to reject the null hypothesis with high probability when the parameters of interest really deviates from its true value. This deviation is denoted as δ. We are interested in detecting the minimum value of δ.

  $1-\beta$ Test Power or Power. The power of a statistical test is the probability that the null hypothesis is rejected when in fact is not true. A common value for power is 0.8, but it depends on the study.

m Total number of observations over time.

σ Standard deviation of the observations. For continuous responses, the quantity $Var(Y_{ij}) = \sigma^2$, i individuals and j measurements over time; represents the unexplained variation in the response. Sometimes it is possible to give an approximation to the real value of σ^2 throughout pilot or previous studies.

R Correlation Structure in Observations (véase A.4). The correlation parameter or parameters can be estimated from previous or pilot studies. We also have to define a correlation structure that better fits our data.

- The formula to estimate sample size in each group is

$$n = \frac{(Z_{1-\alpha/2}+Z_{1-\beta})^2\sigma^2 V}{\delta^2}$$

$$V = \begin{pmatrix} 0 & 1 \end{pmatrix} \left[\begin{pmatrix} 1 & \cdots & 1 \\ t_1 & \cdots & t_m \end{pmatrix} (\sigma^2 R)^{-1} \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_m \end{pmatrix} \right] \begin{pmatrix} 0 \\ 1 \end{pmatrix} [2,2] \tag{5.7}$$

Example

To estimate the sample size using LADES we follow **Statistics** $->$ **Sample Size Estimation Longitudinal Analysis with continuos response**.

The captured data and the main screen of the function are shown in Figure 5.21. For our problem we want to detect a minimum significant difference of $\delta = 4.23$. The standard deviation is obtained from a pilot study conducted before; $\sigma = 13.6$. The probability of rejecting the null hypothesis given that it is true, i.e. type 1 error (α), is 0.05. We will use a power of 0.8. And the alternative hypothesis to be tested is $H_0 : \beta_{1C} \neq \beta_{1T}$ (two-sided).

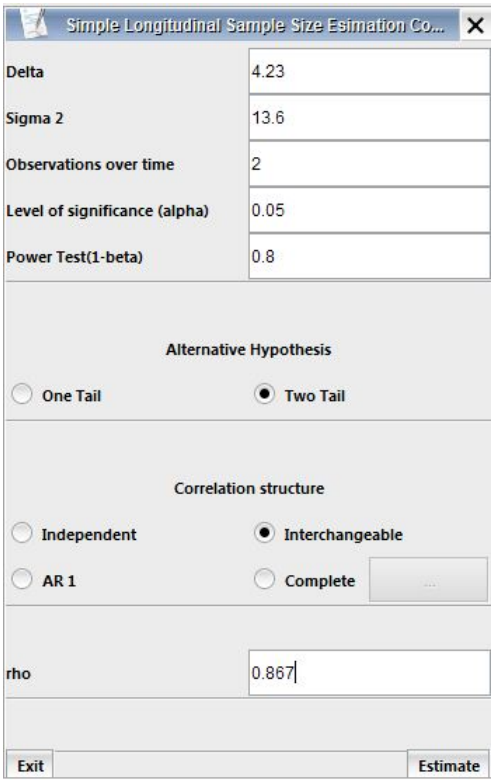


Figure 5.21: Input data Sample Size Estimation, Continuous Response

The more sensible parameter for sample size calculations is the correlation structure. It is assumed that the correlation structure is an *Interchangeable* structure, with $\rho = 0.867$. This means the correlation between observations is the same, 0.867.

The results are shown in Figure 5.22. In conclusion, we will need 44 individuals per group to perform our study.

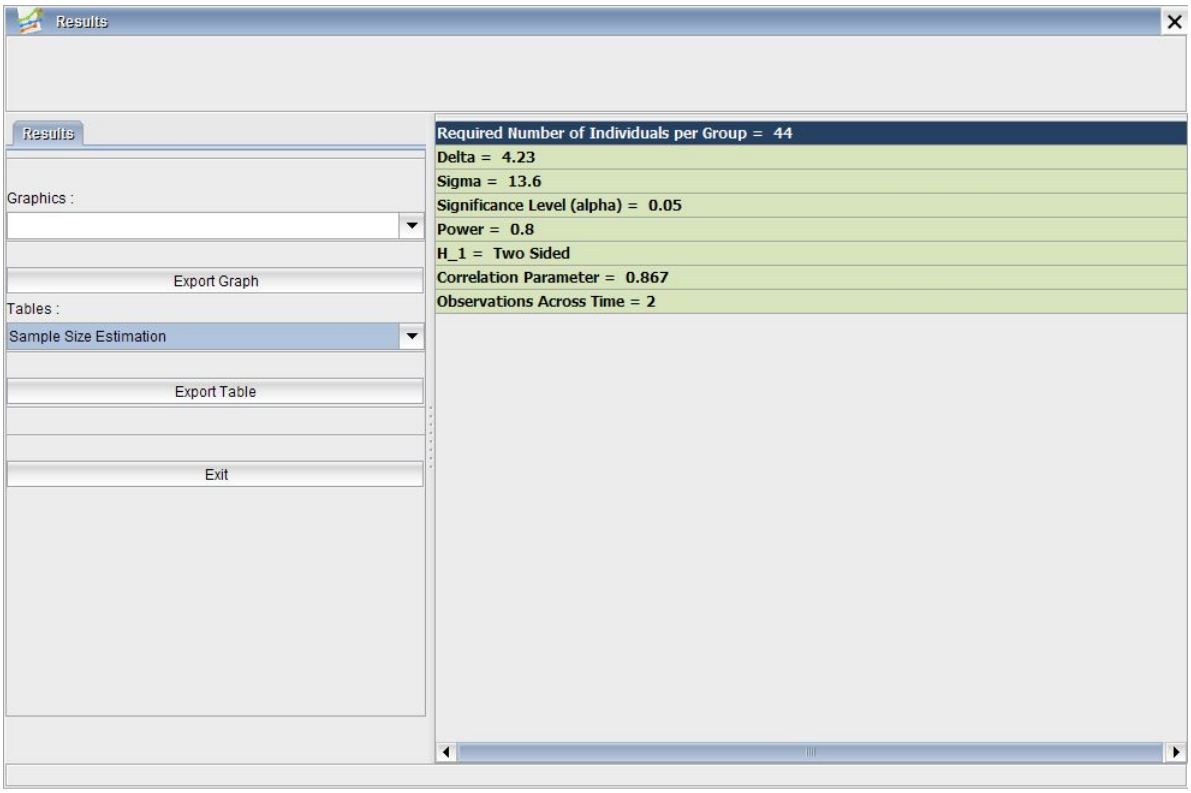


Figure 5.22: Sample size Estimation for continuous response Results

5.4 Sample Size Estimation for Longitudinal Analysis with Binary Response

Before conducting a longitudinal study is necessary to calculate the number of subjects needed to ensure that certain predefined effect is significant. Sample Size calculation is based on assumptions which, if altered, can lead very different sizes. It is related to the significance level of a factor (How many individuals are needed to make some significant effect?).

As an example, we have a pilot study with a binary response variable (1 or 0). They want to test a new drug for patients with stomach ulcers. The response under interest is 1= patient has ulcers, and 0 = no ulcers presented. The duration of treatment is one month, and patients will be reviewed three times. The first review is when the treatment was first administrated. And, patients were evaluated at months 6 and 12. In this study, the performance of this new treatment is compared against a placebo. It is desired that the difference between the proportions is at least 16.5% to consider a significative difference. Previous data collected for the calculation of sample size in this study are shown in Table 5.2.

Information	**Value**
Proportion of patients with no stomach ulcers	0.5
Time Measurements Correlation	0.15
Number of Measurements across time	3
Significance Level	0.05
Test Power	0.8
Significative Difference	16.5 %

Table 5.2: Previous information needed to compute sample size for binary response

Sample size is based on detecting a significant difference between the average change rate of both groups.

Sumary of Sample Size Estimation for binary response

- The null hypothesis on which sample size calculations are based is: $H_0 : \beta_{1C} = \beta_{1T}$. That is, the average change over time for the two groups is the same. The alternative hypothesis can be of two types $H_0 : \beta_{1C} \neq \beta_{1T}$ or $H_0 : \beta_{1C} > \beta_{1T}$ (or, $H_0 : \beta_{1C} < \beta_{1T}$).
- The required information to make the calculations is:
  - α Type 1 Error. This parameter represents the probability the null hypothesis is rejected when it is true. For example, it would correspond to the probability of concluding there is a difference between treatment and control groups when there is not. Usually equal to 0.05.
  - p_1, p_2 Proportion of expected Presents (Y = 1) in group 1 and 2, respectively. Using this two quantities we can define the minimum significance difference to detect, δ. Researchers usually want to reject the null hypothesis with high probability. We are interest in detecting the minimum value of δ.
  - $1-\beta$ Test Power or Power. The power of a statistical test is the probability that the null hypothesis is rejected when it is not true. A common required value for power is at least 0.8, but depends on the study.
  - **m** Total number of observations over time.
  - ρ Correlation Structure in Observations is an *Interchangeable* structure. The parameter ρ defines the complete structure. This correlation pattern for observations can be estimated from previous or pilot studies.
- The formula to estimate sample size is

$$n = \frac{2(Z_{1-\alpha/2} + Z_{1-\beta})^2 \, \bar{p}(1-\bar{p}) \, [1+(m-1)\rho]}{m(p_1 - p_0)^2} \tag{5.8}$$

where $\bar{p} = \frac{p_1+p_2}{2}$.

Example

To estimate the sample size using LADES we follow **Statistics** $->$ **Sample Size Estimation for Longitudinal Analysis with Binary Response**

The captured data and the main screen of the function are shown in Figure 5.23. For our problem we want to detect a minimum significance difference of $\delta = 0.165$ in the proportion of Presents. Thus, we fix both proportions to be $p_1 = 0.5$ (control group) and $p_2 = 0.665$ (treatment group). We use $\alpha = 0.05$, and a power of 0.8. And, the alternative hypothesis to be testes is $H_0 : \beta_{1C} \neq \beta_{1T}$ (two-sided).

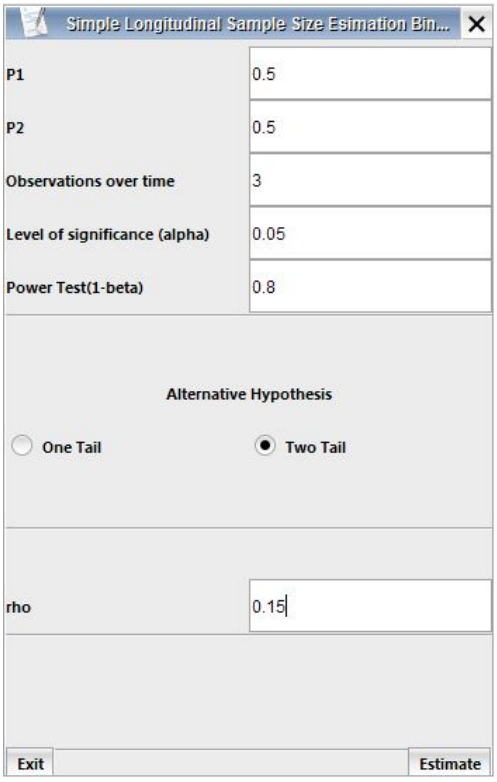


Figure 5.23: Input data Sample Size Estimation, Binary Response

The most sensible parameter for sample size calculation is the correlation structure. It is assumed tbinhat the correlation structure is *Interchangeable* with $\rho = 0.15$. This means the correlation between observations is the same, 0.15.

The results are shown in Figure 5.24. In conclusion, we will need 60 individuals per group to perform our study.

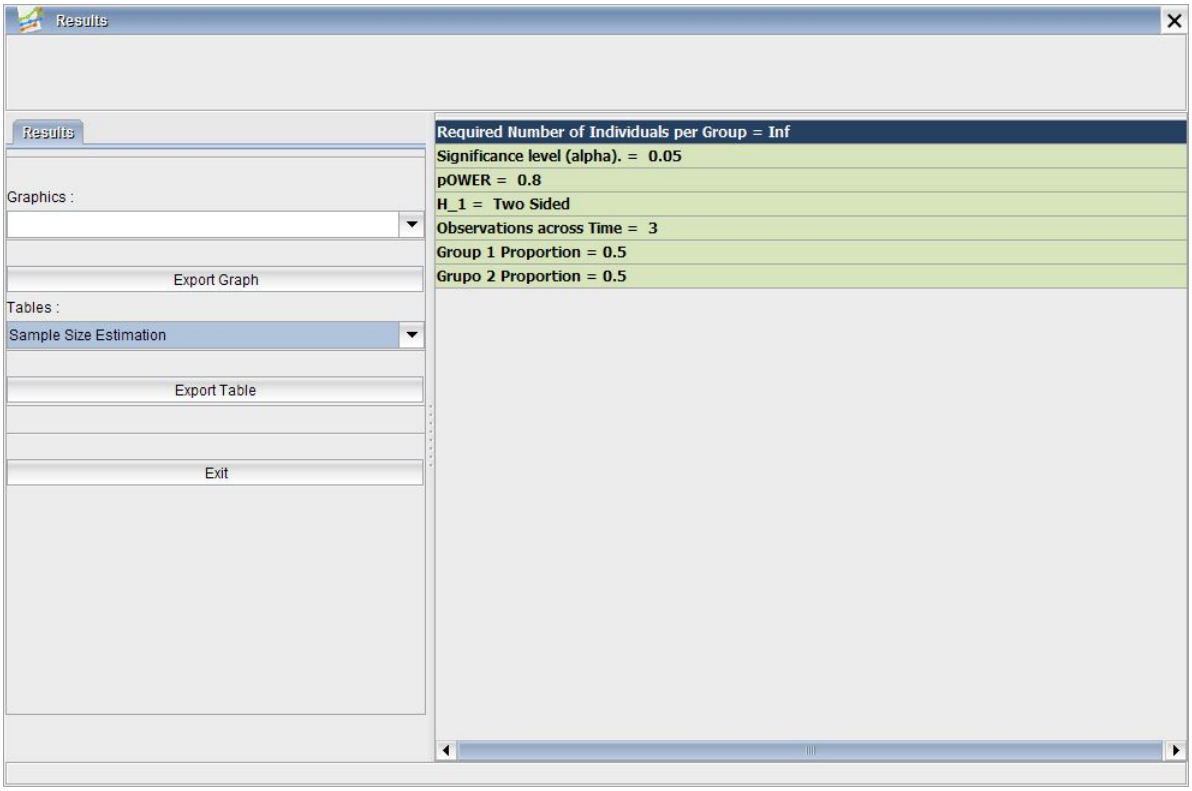
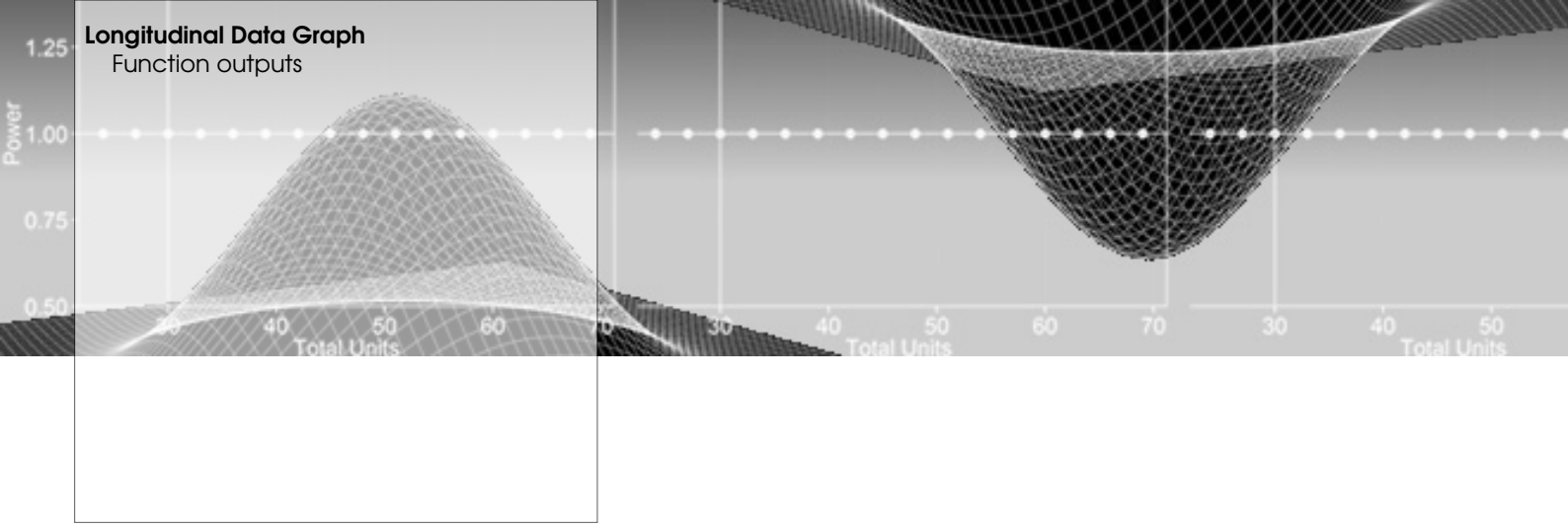


Figure 5.24: Sample size Estimation for binary response Results



6 — Graphics Module

6.1 Longitudinal Data Graph

To explain the use of **Longitudinal Data Graph** function, we will use the data **weights.csv**. After importing the data, we access the function through **Graph** $->$ **Longitudinal Data**. Then, the screen shown in Figure 6.1 appears.

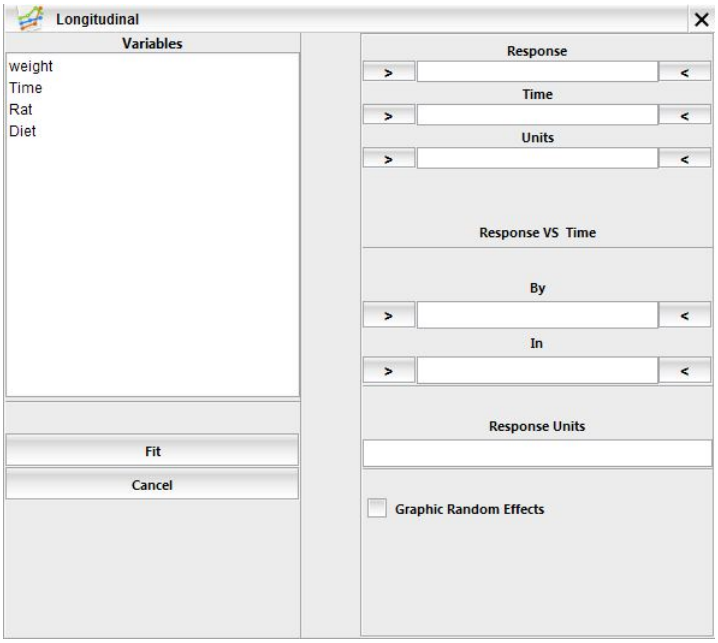


Figure 6.1: Main Screen Longitudinal Data Graph

If we want to plot the response profiles of each individual, we fill only the **Response Time** and **Id** fields; and then click on **Create** (see Figure 6.2). The resulting graph is shown in Figure 6.3.

If now we want to categorize each individual according to some factor of interest; for instance, in our example, we want to analyze the three diet groups. We just have to input the desired factor

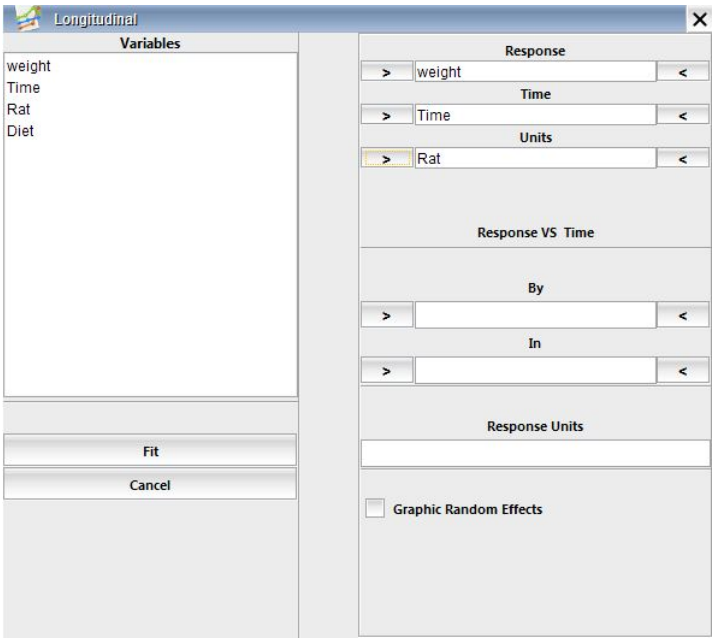


Figure 6.2: How to fill the function Longitudinal Data Graph

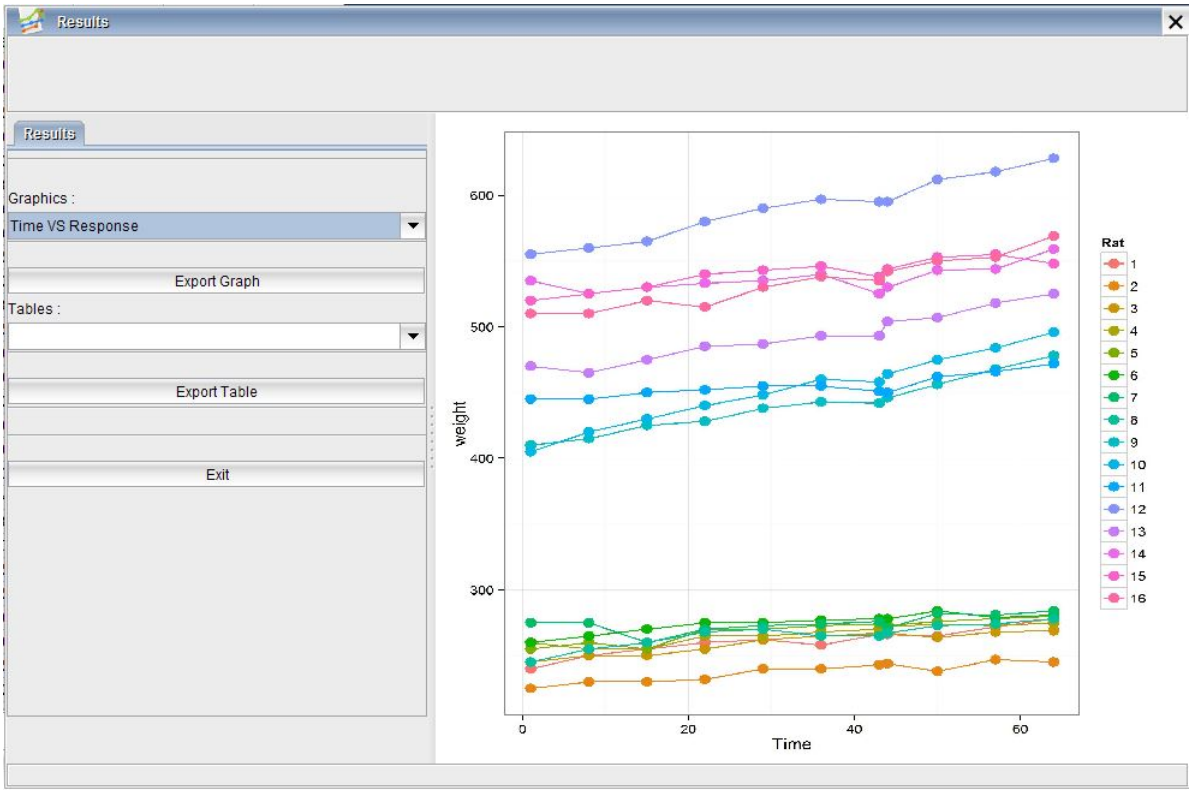


Figure 6.3: Results Simple Longitudinal Graph

Diet in the **By** section, see Figure 6.4. In the resulting display we can have a notion of the effect of each one of the diets under study. We can also add another level of classification factor by placing other factor in the section **In**.

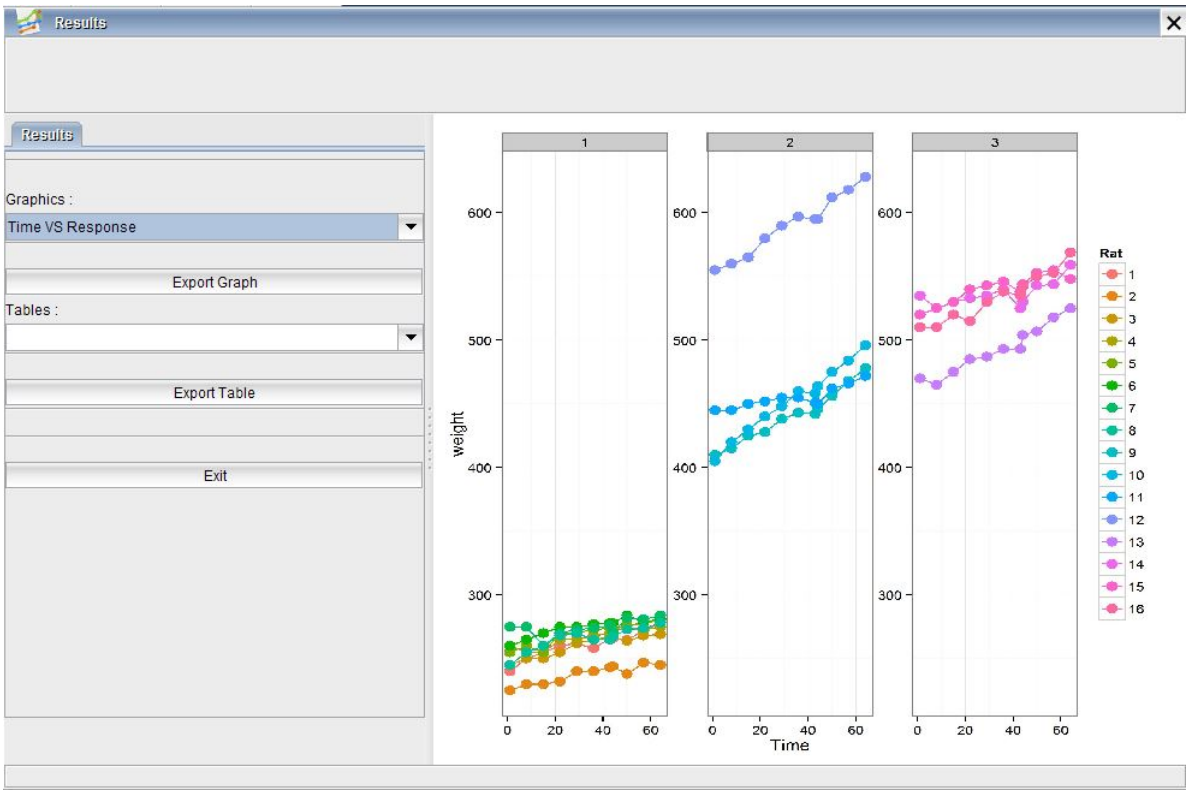


Figure 6.4: Results Longitudinal Plot

We can also include the units of measurement of our observations. In this case, the response variable was the weight of mice in Gr (Figure 6.5). In addition, we choose the option of **Random Effects Plot**, that will be explained soon.

6.1.1 Function outputs

By choosing the **Random Effects Plot** option, a simple linear regression model is fitted to each of the individuals. The regressor used is time.

- **Random Effects Plot.** For the intercept and time estimated coefficients, 95% confidence intervals are constructed which are plotted in Figure 6.6 and 6.7, respectively. This chart allows us identify possible random factors for the intercept and the time factor. If our confidence intervals overlap for all individuals, then we say this factor is constant among individuals. Otherwise, we would need to include a random factor in order to model this difference among individuals.
- **Estimates for Individual Models**. The point estimates of the coefficients for each individual are shown in Figure 6.8.

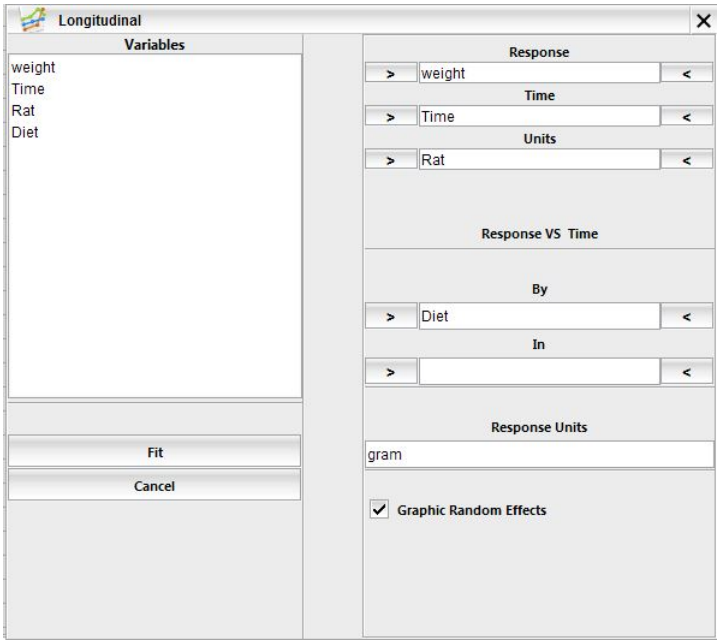


Figure 6.5: Unit inclusion and random effects plot option

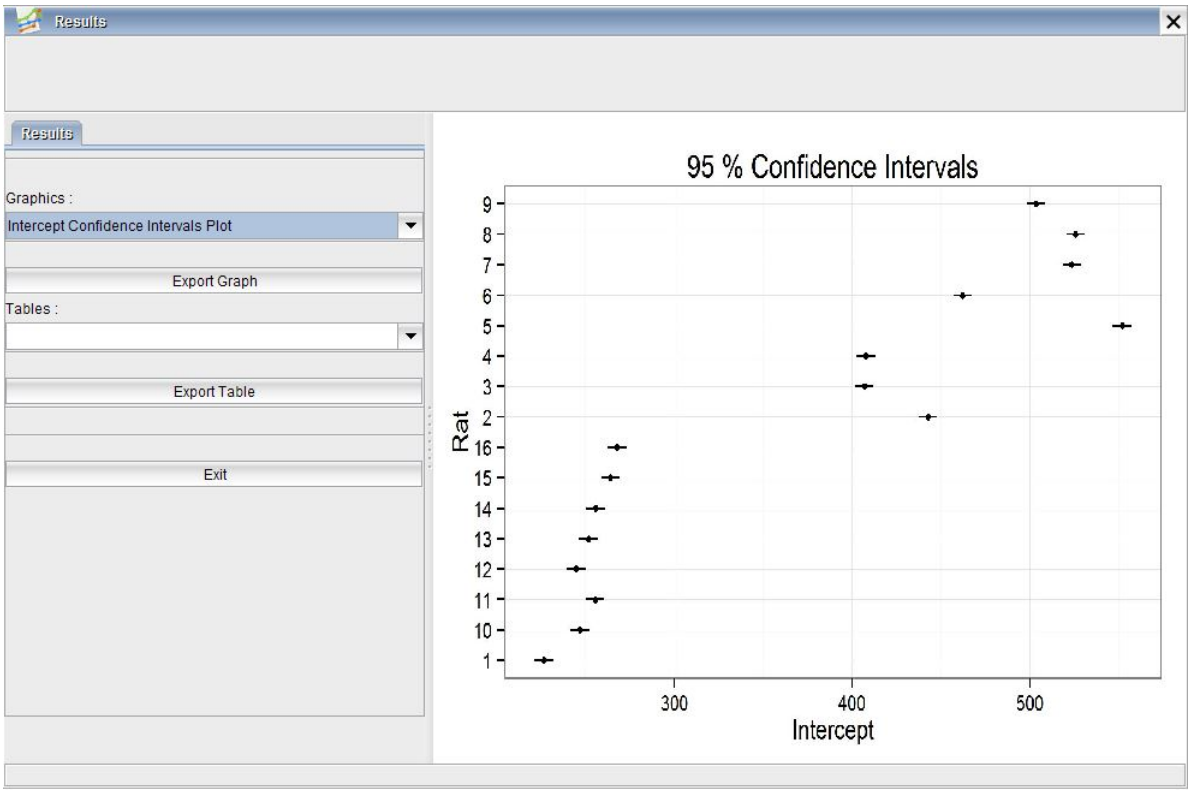


Figure 6.6: Graph of the 95% confidence intervals for intercept coefficients for each individual using a simple linear regression model

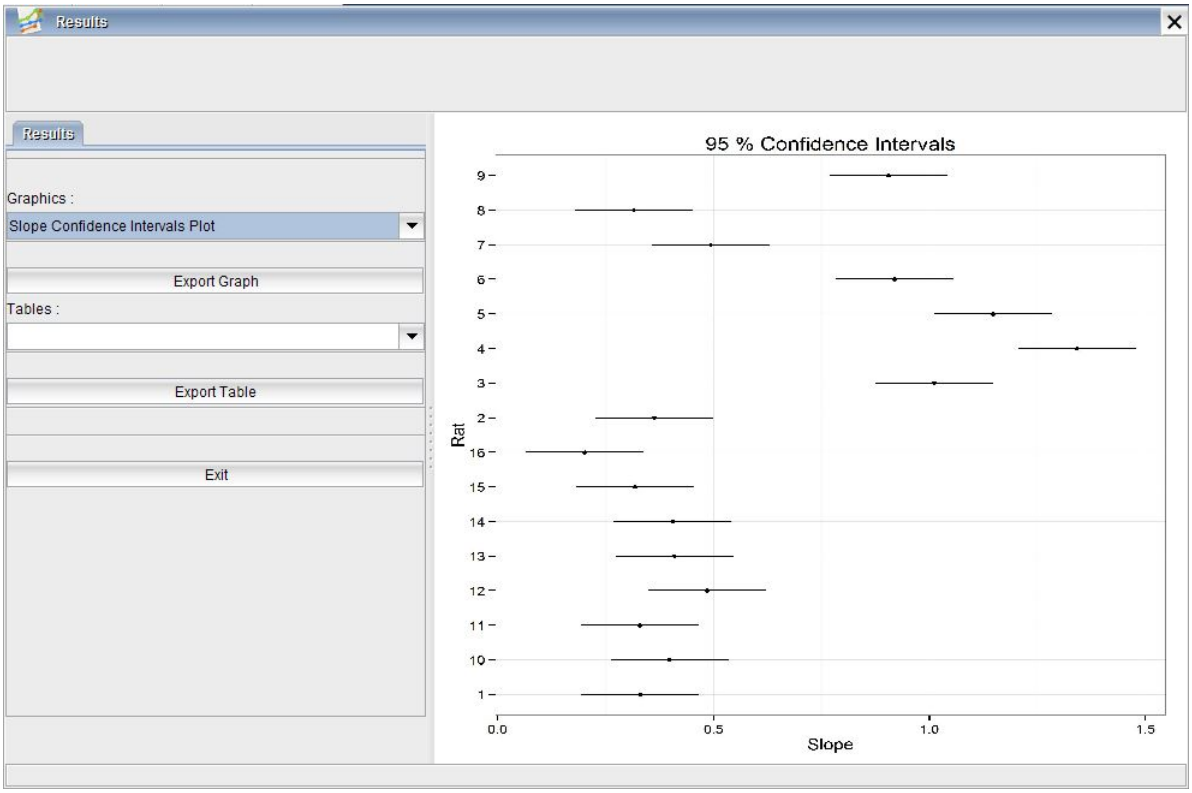
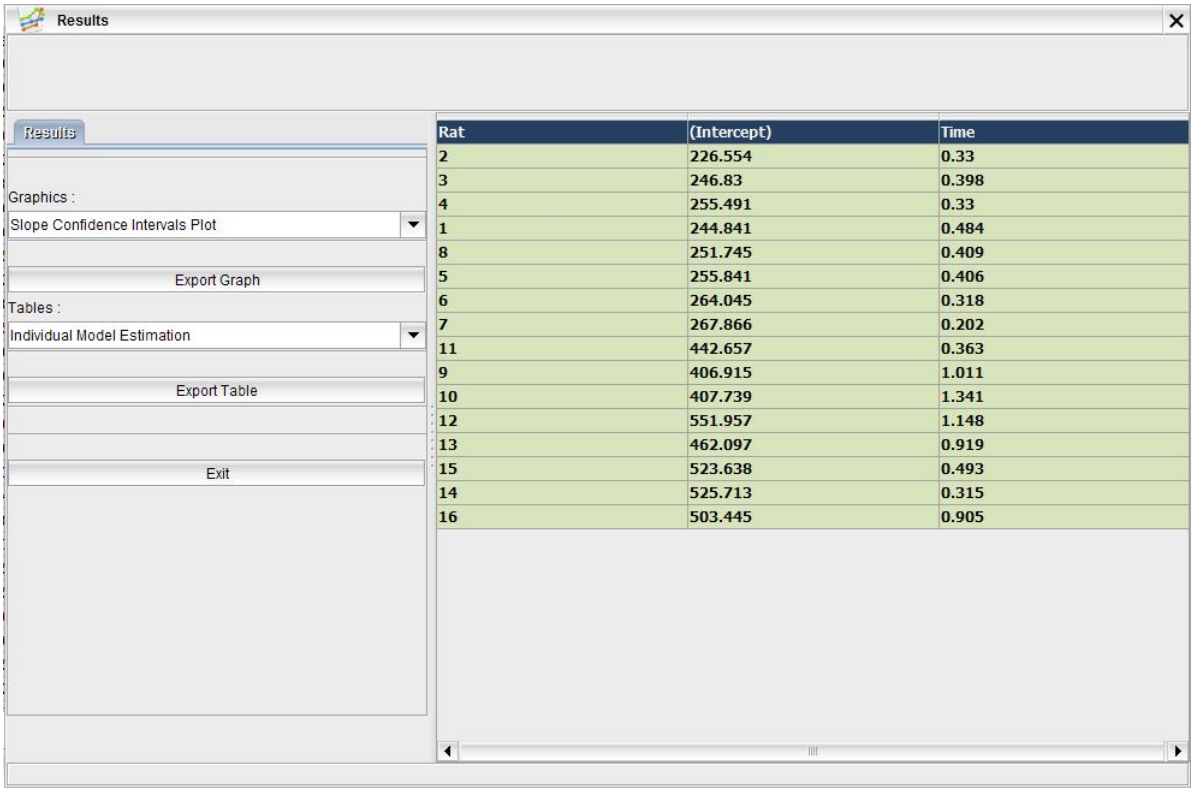
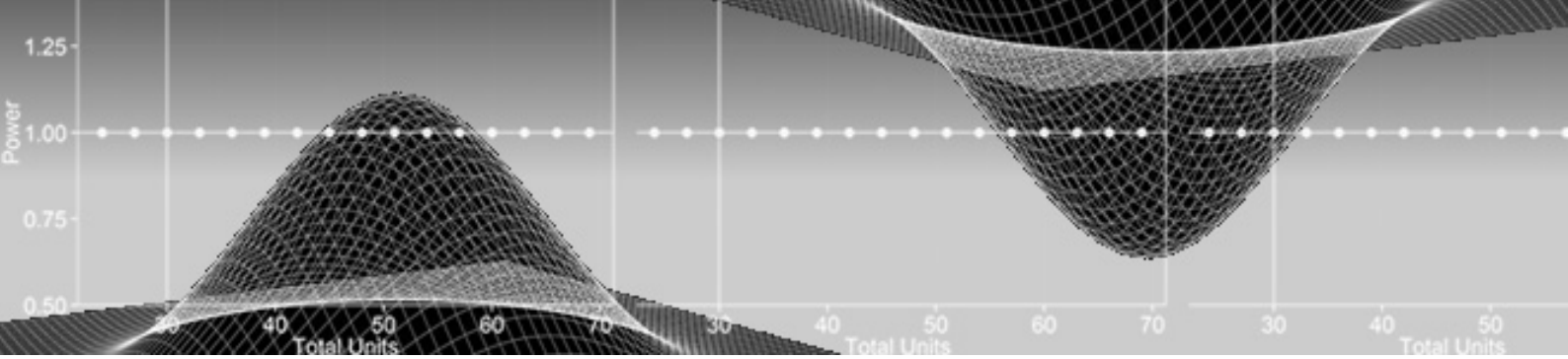


Figure 6.7: Graph of the 95% confidence intervals for time coefficients for each individual using a simple linear regression model



Rat	(Intercept)	Time
2	226.554	0.33
3	246.83	0.398
4	255.491	0.33
1	244.841	0.484
8	251.745	0.409
5	255.841	0.406
6	264.045	0.318
7	267.866	0.202
11	442.657	0.363
9	406.915	1.011
10	407.739	1.341
12	551.957	1.148
13	462.097	0.919
15	523.638	0.493
14	525.713	0.315
16	503.445	0.905

Figure 6.8: Point estimates for the intercept and time coefficients for each individual using a simple linear regression model

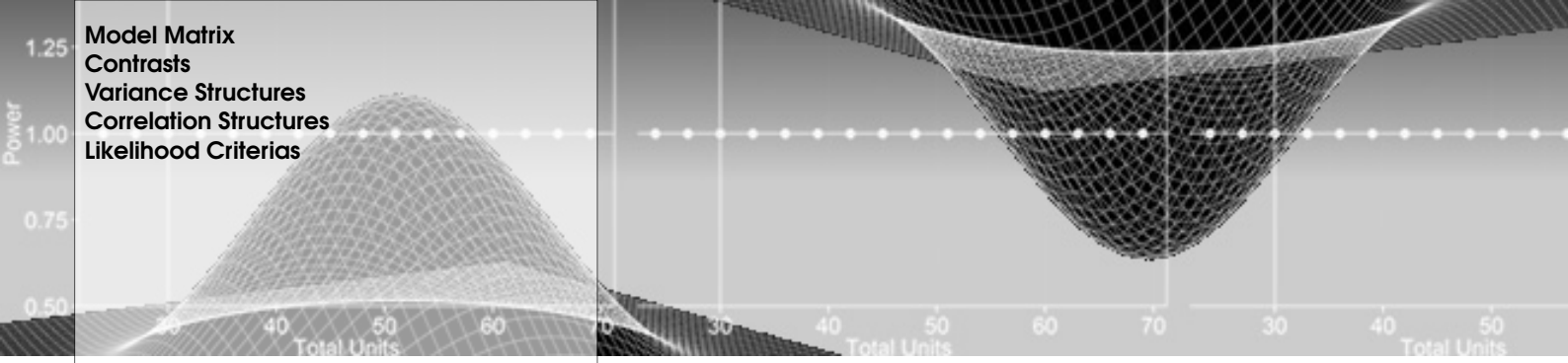


Bibliography

- Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker. lme4: Linear mixed-effects models using Eigen and S4, 2013. R package version 1.0-5.
- P. Diggle. Analysis of Longitudinal Data. Oxford Statistical Science Series. OUP Oxford, 2002.
- J.J. Faraway. Linear Models with R. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2004.
- P. Goos and B. Jones. Optimal Design of Experiments: A Case Study Approach. Wiley, 2011.
- Ulrike Gromping. R package FrF2 for creating and analyzing fractional factorial 2-level designs. Journal of Statistical Software, 56(1):1?56, 2014.
- R W Helms. Intentionally incomplete longitudinal designs: I. methodology and comparison of some full span designs. Stat Med, 11(14- 15):1889?1913, Oct-Nov 1992.
- Søren Højsgaard, Ulrich Halekoh, and Jun Yan. The r package geepack for generalized estimating equations. Journal of Statistical Software, 15/2:1?11, 2006.
- R.A. Johnson and D.W. Wichern. Applied Multivariate Statistical Analysis. Pearson Education International. Pearson Prentice Hall, 2007.
- G.A. Milliken and D.E. Johnson. Analysis of Messy Data: Designed Experiments, Second Edition. Number v. 1. Taylor & Francis, 2009.
- J. Pinheiro and D. Bates. Mixed-Effects Models in S and S-PLUS. Statistics and Computing. Springer New York, 2010.
- Jose Pinheiro, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, and R Core Team. nlme: Linear and Nonlinear Mixed Effects Models, 2013. R package version 3.1-113.
- R Core Team. R: A Language and Environment for Statistical ComputIng. R Foundation for Statistical Computing, Vienna, Austria, 2013.
- Deepayan Sarkar. Lattice: Multivariate Data Visualization with R. Springer, New York, 2008. ISBN 978-0-387-75968-5.
- J.W.R. Twisk. Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide. Cambridge medicine. Cambridge University Press, 2013.
- W. N. Venables and B. D. Ripley. Modern Applied Statistics with S. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.
- Bob Wheeler. AlgDesign: Algorithmic Experimental Design, 2011. R package version

1.1-7.

- Hadley Wickham. ggplot2: elegant graphics for data analysis. Springer New York, 2009.
- C.F.J. Wu and M.S. Hamada. Experiments: Planning, Analysis, and Optimization. Wiley Series in Probability and Statistics. Wiley, 2013.
- Jun Yan. geepack: Yet another package for generalized estimating equations. R-News, 2/3:12?14, 2002.
- Jun Yan and Jason P. Fine. Estimating equations for association structures. Statistics in Medicine, 23:859?880, 2004.
- S L Zeger, K Y Liang, and P S Albert. Models for longitudinal data: a generalized estimating equation approach. Biometrics, 44(4):1049? 1060, Dec 1988.
- A. Zuur, E.N. Ieno, N. Walker, A.A. Saveliev, and G.M. Smith. Mixed Effects Models and Extensions in Ecology with R. Statistics for biology and health. Springer, 2009.



A — Term Definitions

A.1 Model Matrix

If N observations are collected in an experiment, and the linear model is

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \varepsilon_i \quad i = 1, \ldots, N \tag{A.1}$$

where y_i represents i-th response, and $x_{i1}, \ldots, x_{ik}$ are the corresponding values of the k dependent variables. These N equations can be written in matrix form as

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{A.2}$$

where $\mathbf{y} = (y_1, \ldots, y_N)^T$ is the response vector of dimensions $N \times 1$, $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$ is the regression coefficient vector, $(k+1) \times 1$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_N)^T$ is the error vector $N \times 1$; and $\mathbf{X}$, is the **model matrix** of dimensions $N \times (k+1)$, given as

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \ldots & x_{Nk} \end{pmatrix} \tag{A.3}$$

A.2 Contrasts

Treatment

Consider a qualitative factor having four levels. This factor can be encoded using three 'dummy' variables. This contrasts matrix describes the codification where the columns represent 'dummy' variables and the rows represent levels:

$$\begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{A.4}$$

In this encoding, the level one (first row) is considered as the reference level at which the other levels will be compared. This is similar to the implementation of a control group. Then, if it exists, it is recommended to assign it to level one.

As a result, the parameter for the dummy variables (regression coefficient; columns of matrix A.4) represents the difference between certain level and the reference level.

Helmert

The codification for this type of contrast is

$$\begin{pmatrix} -1 & -1 & -1 \\ 1 & -1 & -1 \\ 0 & 2 & -1 \\ 0 & 0 & 3 \end{pmatrix}$$

If you have an equal number of observations at each level (ie, a balanced design) then the dummy variables will be orthogonal to each other and to the intercept. This codification is not very suitable for the interpretation of the results, but can be used in some special cases. For our qualitative variable, the parameter of the second dummy variable (coefficient regression; first column) refers to the average difference between level 1 and level 2. The interpretation of the parameter of the third dummy variable relates to the average difference between level 1 and level 2 with twice the average of observations at level 3. The same holds for the fourth dummy variable.

A.3 Variance Structures

Fixed

This structure represents a variance function without parameters, but with one variable affecting the variance. This structure is used when the variance within-groups is defined by a proportional constant. This structure is used, for example, when we have evidence that the variance within-group increases linearly over time.

$$Var(\varepsilon_{ij}) = \sigma^2 t_{ij} \tag{A.5}$$

i individuals, j measures in time.

Ident

This structure represents a variance model in which the variances for each level of stratification of the s variables is different. Thus, different values are taken in the set 1, 2, ..., S, that is

$$Var(\varepsilon_{ij}) = \sigma^2 \delta^2_{S_{ij}} \tag{A.6}$$

The variance model (A.3) uses $S+1$ parameters to represent the S variances.

Power

The variance model for this structure is

$$Var(\varepsilon_{ij}) = \sigma^2 |v_{ij}|^{2\delta} \tag{A.7}$$

which represents a potential absolute change in variance by variable. The parameter δ is not restricted, then (A.3) can be used in cases where the variance increases or decreases with the absolute value of a variable.

Exponencial

This variance model is represented by the structure

$$Var(\varepsilon_{ij}) = \sigma^2 (2\delta v_{ij}) \tag{A.8}$$

It can be explained as an exponential function of the variance through a variable. The parameter δ is not restricted, then (A.3) can be modeled in cases where the variance increases or decreases with a variable.

A.4 Correlation Structures

Exchangeable

This is the simplest correlation structure, which assumes equal correlations between all errors within-groups belonging to the same group. An example of this correlation structure can be seen in the following matrix where we considered 4 observations over time:

$$\begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix} \tag{A.9}$$

When using this structure to represent the relationship between data taken across time, it is only necessary to estimate a parameter ρ.

General (Full)

This structure represents the most complex structure. Each correlation in the data is represented by a different parameter, the corresponding correlation structure is as follows:

$$\begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_4 & \rho_5 \\ \rho_2 & \rho_4 & 1 & \rho_6 \\ \rho_3 & \rho_5 & \rho_6 & 1 \end{pmatrix} \tag{A.10}$$

Because the number of parameters in (A.4) is increased quadratically with the number of observations of individuals over time, the correlation structure usually lead us to an overparameterized model. When there are few observations per group, general correlation structure is a useful tool to model the data.

Autoregressive order 1, AR1

The AR (1) model is the simplest of the autoregressive models. The correlation structure using this model represents a decrease in the correlation as that observations are more and more separated. An example of this structure can be observed below:

$$\begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix} \tag{A.11}$$

We note using structure only one parameter has to be estimated, ρ. This matrix is the most used when fitting a longitudinal analysis model since, in most of the cases, more distant observations tend to be less correlated. Moreover, the number of parameters to estimate is just one.

Independent This structure is used when there is no relationship between observations. An example of this structure is:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{A.12}$$

It is assumed that when we are working with the linear regression model, our errors follow this structure. Independent correlation matrix does not help us much when we want to use longitudinal data analysis; where the interest is to model the correlation of measurements over time.

A.5 Likelihood Criterias

Akaike Information Criteria Criterion (AIC) and the *Bayesian Information Criterion* (BIC) are evaluated as follows:

$$AIC = -2\log Lik + 2n_{par} \tag{A.13}$$

$$BIC = -2\log Lik + n_{par}\log(N) \tag{A.14}$$

where n_{par} denotes the number of parameters in the model, and N is the total number of observations used to fit the model. Under these definitions, "better is better". So if we are using AIC to compare two or more models for the same data, we would prefer the model with the lowest AIC. Similarly, when using BIC prefer the model with the lowest BIC.

As already mentioned, what is desired is to minimize AIC or BIC. Models that consider many factors tend to fit the data better, but require more parameters. Then the best choice for our model is to a balance between between its fitting and its number of parameters. BIC severely penalizes models with more parameters, then tends to prefer small models in comparison with the AIC.